July 2017

# Information Metrics for Predictive Modeling and Machine Learning

Kostantinos Gourgoulias

# INFORMATION METRICS FOR PREDICTIVE MODELING AND MACHINE LEARNING

A Dissertation Presented

by

KONSTANTINOS GOURGOULIAS

Submitted to the Graduate School of the
University of Massachusetts Amherst in partial fulfillment
of the requirements for the degree of

DOCTOR OF PHILOSOPHY

May 2017

Department of Mathematics and Statistics

# INFORMATION METRICS FOR PREDICTIVE MODELING AND MACHINE LEARNING

A Dissertation Presented

by

KONSTANTINOS GOURGOULIAS

Approved as to style and content by:

Markos A. Katsoulakis, Chair

Luc Rey-Bellet, Member

Patrick Flaherty, Member

Arya Mazumdar, Member

Farshid Hajir, Department Chair
Department of Mathematics and Statistics

# DEDICATION

*To my parents, Vaggeli and Eirini, and my sister, Maria.*

# ACKNOWLEDGMENTS

I'm a firm believer that one's creative work is heavily influenced by the social and professional interactions that are experienced. In this regard, I have to start by thanking my first advisor, Professor Katsoulakis. His advice was—and still is—always on point and he was able to adapt to my style of research, letting me search for ideas on my own but also keeping me focused on the main goal. I have to also thank Professor Rey-Bellet who joined as a co-advisor soon after. It has been incredible luck to be able to work under both and I have learned a lot from doing so!

The presentation of the results has been greatly improved by comments from the rest of my committee: Professor Flaherty and Professor Mazumdar. Thank you both for your input on my work and your unique perspectives!

I would also like to thank the Department of Mathematics and Statistics in general, for their amazing support over the years with the various issues that popped up. Thanks to the people of RCF, especially Ken Pollard, I was able to work efficiently and finish writing this thesis on time. Big thanks also to Professor Johnston for allowing me to think about a new project during my second year and his support throughout the program.

A big part of the support came from all my friends in New England and beyond. I can't possibly fit all their names here, but I would especially like to thank: Aaron, Ankita, Betsy, Chucky and Kate, Dom and Joy, Emily and Sean, Giorgos, Jezabel, Jie, Katerina, Marina, Mark, Matt, Michael, Panos, Rodrigo, Shermin, Steele and Stephen. Also, thanks go to every person associated with GRiD.

Finally, I have to thank my partner, Ria, for being patient with all my ideas and pragmatic enough to tell me which ones are worth pursuing.

# ABSTRACT

## INFORMATION METRICS FOR PREDICTIVE MODELING AND MACHINE LEARNING

MAY 2017

KONSTANTINOS GOURGOULIAS

B.Sc., UNIVERSITY OF CRETE

Ph.D., UNIVERSITY OF MASSACHUSETTS AMHERST

Directed by: Professor Markos A. Katsoulakis

The ever-increasing complexity of the models used in predictive modeling and data science and their use for prediction and inference has made the development of tools for uncertainty quantification and model selection especially important. In this work, we seek to understand the various trade-offs associated with the simulation of stochastic systems. Some trade-offs are computational, e.g., execution time of an algorithm versus accuracy of simulation. Others are analytical: whether or not we are able to find tractable substitutes for quantities of interest, e.g., distributions, ergodic averages, etc.

The first two chapters of this thesis deal with the study of the long-time behavior of parallel lattice Kinetic Monte Carlo (PL-KMC) algorithms for interacting particle systems. We introduce the relative entropy rate (RER) as a measure of long-time loss of information and illustrate that it is a computable *a posteriori* quantity. The RER

can act as an information criterion (IC), discriminating between different parameter choices for the schemes and allowing comparisons at equilibrium. We make explicit how the RER scales with the time-step and the size of the system and that it captures details about the connectivity of the original process.

Another feature of long-time behavior is time-reversibility, which some physical systems naturally exhibit. Unfortunately, due to the domain and time-discretization, PL-KMC cannot conserve this property. To quantify the loss of reversibility, we introduce the entropy production rate (EPR) as an IC for comparisons between different schemes. We show that the EPR shares a lot of the properties of the RER and can be estimated efficiently from data.

The last chapter discusses uncertainty quantification for model bias. By connecting a recently derived goal-oriented divergence and concentration bounds, we define new divergences that provide computable bounds for model bias. The new bounds scale appropriately with data and become progressively more accurate depending on available information about the models and the quantities of interest. We discuss how the bounds allow us to bypass computationally expensive Monte Carlo sampling or specialized methods, e.g., Multilevel Monte Carlo.

# TABLE OF CONTENTS

# LIST OF TABLES

**Table** **Page**

# LIST OF FIGURES

**Figure** **Page**

# INTRODUCTION

The main topic of this thesis is the development of information metrics for uncertainty quantification, numerical analysis, and model selection for stochastic systems. At a high level, we seek to understand the various trade-offs associated with the simulation of stochastic systems through techniques from information theory and model selection. Some trade-offs are computational, e.g., execution time of an algorithm versus accuracy of simulation. Others are analytical: whether or not we are able to find tractable substitutes for quantities of interest, e.g., distributions, ergodic averages, etc.

The first two chapters of the thesis deal with the study of the long-time behavior of parallel schemes used for the simulation of lattice dynamics. Then, the last chapter discusses tractable information metrics for model bias.

**Information criteria for Parallel Kinetic Monte Carlo:**

Schemes that depend on operator splitting have found wide applicability within the domain of simulation of complex chemical reaction systems, biological systems, interacting particle systems,etc. However, such schemes were first used in the numerical solution of differential equations [52].

**Example 1** (Operator Splitting for differential equations)**.** *Consider a bounded operator $L$ and the differential equation:*

$$u' = Lu. \tag{1}$$

*For example, if $L$ was a square matrix,* (1) *would correspond to a system of first-order differential equations. Regardless, we can express the solution of* (1) *in the language of operator semigroups [54] as*

$$u(t) = e^{\Delta t L} u(0). \tag{2}$$

*Next, we assume that there exist bounded operators $L_1$ such that $L = L_1 + L_2$. If $L$ was a square matrix, then for any matrix $C$ of the same dimensions, we have the decomposition with $L_1 = C$ and $L_2 = L - L_1$. Now that we have split $L$ into $L_1$ and $L_2$, we can built approximations to the solution of* (1) *in the following way. First, we notice that:*

$$\begin{aligned} u' = L_1 u &\Rightarrow u(\Delta t) = e^{\Delta t L_1} u(0), \\ u' = L_2 u &\Rightarrow u(\Delta t) = e^{\Delta t L_2} u(0). \end{aligned} \tag{3}$$

*Then, by the Trotter product formula [60]:*

$$e^{\Delta t L} = \lim_{n \to \infty} (e^{\Delta t / n L_1} e^{\Delta t / n L_2})^n.$$

*This motivates approximations of $e^{\Delta t L} u(0)$ by products of $e^{\Delta t L_1}$ and $e^{\Delta t L_2}$. A particular approximation is the Lie splitting, $e^{\Delta t L_1} e^{\Delta t L_2} u(0)$. In fact, as $L_1, L_2$ are bounded operators, we can use the series expansions of $e^{\Delta t L}$ and $e^{\Delta t L_1} e^{\Delta t L_2}$ to explicitly calculate the error of the approximation:*

$$\begin{aligned} e^{\Delta t L} u(0) &= \left(1 + \Delta t (L_1 + L_2) + \frac{\Delta t^2}{2} (L_1 + L_2)^2 + o(\Delta t^2)\right) u(0), \\ e^{\Delta t L_1} e^{\Delta t L_2} u(0) &= \left(1 + \Delta t (L_1 + L_2) + \frac{\Delta t^2}{2} \left(L_1^2 + L_2^2 + 2 L_1 L_2\right) + o(\Delta t^2)\right) u(0), \\ e^{\Delta t L} u(0) - e^{\Delta t L_1} e^{\Delta t L_2} u(0) &= \frac{\Delta t^2}{2} [L_2, L_1] u(0) + o(\Delta t^2). \end{aligned}$$

*where* $[L_1, L_2] = L_1L_2 - L_2L_1$ *is the Lie bracket of* $L_1, L_2$*. Note that if* $[L_1, L_2] = 0$*, then the splitting method we have used is exact.*

Example 1 showcases an operator splitting method motivated from the structure of the operator $L$. In this work, we will consider operator splittings that are imposed by a domain decomposition method, like the checkerboard decomposition in Figure 1. We will see in Section 1.1.2 that $[A, B]$ captures the error of the schemes we will study and has a geometric interpretation.

The recipe of splitting the system into components that can be simulated separately has led to more efficient algorithms, sometimes because some of the components can be solved explicitly (such as in Example 1 or, in chemical reaction systems [33]) and sometimes because the splitting allows for parallel computations [5, 56]. In parallel with the development of those algorithms, there has also been a growing amount of work toward the numerical analysis of splitting methods for stochastic dynamics in different contexts [33, 5, 4, 30, 24, 6]. For parallel lattice kinetic Monte Carlo (PL-KMC), the authors in [5] developed a general framework that connects lattice decompositions to operator splitting. Then, in [4], error estimates were provided for bounded time intervals along with comparisons between different splitting schemes. One of the important contributions of their work was to highlight the connection of

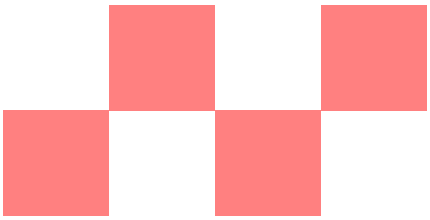

**Figure 1.** A checkerboard decomposition of a rectangular lattice.An operator $L$ that acts on the whole lattice will be decomposed to $L_1+L_2$, where $L_1$ only takes the white sub-lattices into consideration and $L_2$ which only takes the red. Every decomposition of the lattice, e.g., blocks, stripes, etc., corresponds to a different operator splitting of $L$; see Section 1.1.2.

the error with the commutator associated with the splitting and how it affects the efficiency of the scheme.

Classical numerical analysis techniques, such as the study of the local error of the splitting scheme and expansions of the global error [59], work well in providing error estimates for bounded time intervals. However those approaches become non-discriminating when we are interested in long-time results. So, if we wish to sample from an equilibrated system, we have to carefully quantify the errors generated by the scheme at that regime. Approaches for tackling this problem are varied: In the case of SDEs, study of the long-time behavior has been done by employing Poisson equations [51]. For Lie–Trotter splittings, backward error analysis [1] has been used to study the performance of the schemes in capturing the stationary distribution when simulating Langevin dynamics (but see also [45]).

In this work, we introduce the relative entropy on path space per unit time, also known as relative entropy rate (RER), as a tool that can capture the long-time loss of information when using splitting schemes for PL-KMC; see Chapter 1. Through rigorous asymptotics, shown in Section 1.4, we provide an *a posteriori* error expansion of RER in terms of $\Delta t$ and connect RER with quantities central to the classical numerical analysis of splitting schemes, like the commutator and the order of the local error of the splitting method. After deriving computable estimators from our a posteriori expansions for the highest-order term coefficients, we estimate them with the use of SPPARKS [56], a parallel KMC simulator, and use them to compare two well-known splitting schemes, the Lie and Strang splittings. Also, we illustrate how a practitioner can use the RER as an information criterion for selecting schemes that takes into account both long-time accuracy and communication cost. We prove that the RER captures how much the parallel operating splitting scheme preserves the connectivity of the serial algorithm; see Theorem 15 for details.

**Quantifying the loss of time-reversibility with information metrics:**

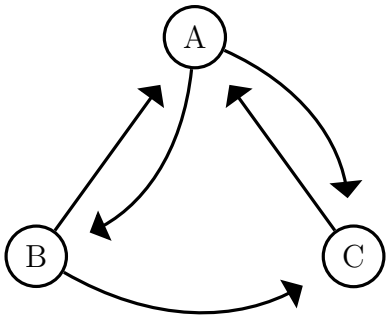


**Figure 2.** A graph illustrating the connectivity between different states of a Markov chain — an arrow from one state to another means that the Markov chain can make that jump in one step. A parallel algorithm that respects this connectivity will be more accurate at long times; see Section 1.7 and in particular the simple Markov chain example in Section 1.7.1.

Another aspect of long-time behavior concerns systems with time-reversible dynamics. That symmetry is often an integral part of the physical structure of the model, for example in the simulation of interacting diffusions or adsorption/desorption mechanisms. While in such cases the time-reversal symmetry is preserved under the serial KMC simulation (typically by enforcing the detailed balance condition), the time-discretization, domain decomposition, and breakdown of serial communication of the parallelized algorithm may lead to loss of detailed balance, and thus of reversibility. There exists some literature on constructing parallel algorithms that preserve the detailed balance (DB) condition [53]. In those algorithms, the scheme picks a schedule for sweeping over the lattice sub-domains, executes it by simulating each sub-domain forward in time for a fixed number of time steps according to the schedule, and then picks a new schedule. For the adjustment to the correct timescale, computation of an equilibrium autocorrelation function is also required. Although such schemes resemble the random Lie-Trotter splitting [4] and they can be numerically analyzed in a similar manner, we will not discuss them here, mainly due to the technical differences with schemes that employ a fixed computational schedule [5].

In [35] the authors used the entropy production rate (EPR) as an information metric to quantify the loss of reversibility for the Euler-Maruyama and Milstein schemes for stochastic differential equations (SDEs), as well as BBK schemes for Langevin dynamics. This idea was motivated by concepts in non-equilibrium statistical mechanics, originally developed to understand the long-time dynamics and the fluctuations in non-equilibrium steady states [49, 50, 42, 41, 25]. Then, in [35], the same methods were used as numerical tools to assess the loss of reversibility of numerical schemes for SDEs. More specifically, they computed the EPR with the Gallavotti-Cohen action functional [42] as an estimator for different numerical schemes. It was demonstrated that the scheme performance in controlling the loss of reversibility can vary greatly. In particular, the Euler-Maruyama scheme for SDEs with multiplicative noise can break reversibility in an unrecoverable manner regardless of the size of the time step [35, Theorem 3.7].

We apply a similar perspective for the study of splitting schemes in parallel KMC (discussed in Chapter 2). In sharp contrast with the schemes for SDEs, for the class of systems we can simulate with PL-KMC, the transition probabilities are intractable to compute or not available at all. Because of this, a new approach is required, which is why we express the EPR as an asymptotic expansion in the scheme's time step by using the semigroup theory for Markov chains. We demonstrate that the coefficients of the expansion of the EPR depend on the transition rates of the model and, most importantly, can be estimated as ergodic averages by samples from the parallel algorithm. We also show that the required computations for the estimation of the coefficients scale with the size of the boundary between sub-domains on the lattice in a manner that depends on the scheme selected. Therefore, by appropriate normalization, we can calculate the entropy production rate per lattice site, i.e. independent of system size. We thus obtain an *a posteriori* expansion for the estimator of the EPR, which can be used as a diagnostic tool that can be calculated on a system of smaller

size than the targeted one, and/or even ran with a simple serial implementation of the parallel algorithm.

**Guarantees for model bias via concentration bounds:**

The last part of the thesis is concerned with the quantification of model bias. That is, given probabilistic models $P, Q$ and a quantity of interest (QoI), $f$, we are interested in guarantees about the model bias,

$$F_{-}(Q, P, f) \leq \mathbb{E}_Q[f] - \mathbb{E}_P[f] \leq F_{+}(Q, P, f), \tag{4}$$

where $F_{\pm}(Q, P, f)$ incorporate attributes/data about the models and the QoI.

Such guarantees are necessitated if we wish to carry out careful prediction and inference when, for example, parts of the model $P$ are unknown or we have a finite sampling budget from $P$ — this is the case with posterior distributions in Bayesian inference. To avoid expensive Monte Carlo sampling, practitioners turn to tractable surrogates of $P$. Those are models $Q$ that approximate $P$ and are simple to sample or evaluate; see variational inference from [9]. Once such a model $Q$ is available, we can use it to estimate descriptive statistics, e.g., mean, variance, cumulative distribution functions, etc. Note that this point of view encompasses PL-KMC as well; $P$ would be the description of the serial Markov process, that is expensive to sample from, and $Q$ describes a parallel scheme.

The point behind building such a surrogate $Q$ rests in trading accuracy for computational time. However, when guarantees for model bias are not present, $Q$ can end up being an erroneous description of our data and we won't know even after we evaluate our QoI. A particular example comes from *algorithmic / machine learning bias*. *Pro Publica*, a non-profit investigative journalism organization studied the risk scores associated to more than 7000 people arrested in Broward County, FL, between 2013 and 2014, to see if offenders with high risk score ended up committing more crimes

over the next two years. The model was designed by Northpointe and the algorithm is put to use within the U.S. criminal justice system. However, after analyzing the data and the risk scores, *Pro Publica* showed that they were only 20% accurate in predicting violent crimes happening within the next two years — the algorithm was unfairly biased towards black defendants, giving them twice the rate of white defendants. As predictive modeling and machine learning integrate more with other fields of study, more examples of model bias will also be revealed, for example, affecting credit offers, who gets to see specific online job listings, etc.; see [40]

Model bias that is inherent in the process of building a surrogate $Q$ comes from variational inference (VI); see [48, 55, 32]. For probability distributions $P, Q$, with $Q$ absolutely continuous with respect to $P$, the Kullback-Leibler divergence (KL) is defined as:

$$R(Q\|P) = \int \log \frac{dQ}{dP} dQ,$$

with $dQ/dP$ being the associated Radon-Nikodym derivative. As the KL is a divergence, $R(Q\|P) = 0$ is equivalent to $Q = P$. Now, minimizing the KL over a family of tractable distributions $Q \in \mathcal{Q}$ allows us to find approximations to $P$ that are simple to evaluate or sample. However, this also implies that $R(Q\|P) \neq 0$ for all $Q \in \mathcal{Q}$, which leads to model bias. A *controlled* increase in model bias is acceptable, as it allows us to conduct inference without the potentially computationally-expensive sampling of Markov Chain Monte Carlo (MCMC). Other applications that make use of similar approximations lie in machine learning [63, 10] and coarse-graining [14, 34, 58, 7, 8].

When using information metrics, a particular challenge is to translate the uncertainty captured by the object, e.g., KL, $\chi^2$ divergence, Hellinger distance, etc., to bias of an observable. A somewhat effective way to accomplish this is through the use of information inequalities, i.e., bounds of the bias in terms of a divergence. Given a

QoI $f$ and probability distributions $P, Q$, such that $Q$ is absolutely continuous with respect to $P$, the Csiszar-Kullback-Pinsker (CKP) inequality [61] states:

$$|\mathbb{E}_Q[f] - \mathbb{E}_P[f]| \leq \|f\|_\infty \sqrt{2R(Q\|P)}, \tag{5}$$

Various other inequalities like (5) exist, such as the Le Cam bound [61] that involves the Hellinger distance (but also see [19] for a tighter bound involving the Hellinger) and the Chapman-Robins [43] that depends on the $\chi^2$ divergence of $Q$ with respect to $P$.

The bound in (5) scales with the magnitude of the QoI $f$ and thus the right-hand side has the potential to become non-discriminating for the bias—even in situations where the bias is small! In fact, a variety of information bounds—including the ones referenced above—were studied in [37] and were found to be non-scalable for high-dimensional systems. Then, in [21] and [16], a pair of *goal-oriented divergences* (GO) was derived that take into account both the uncertainty of the observable *and* the value of the KL divergence between the models compared:

$$\begin{aligned} \Xi_\pm(Q\|P; f) &:= \inf_{c>0} \left\{ \frac{1}{c} \log M_P(\pm c; \tilde{f}) + \frac{1}{c} R(Q\|P) \right\}, \\ M_P(c; \tilde{f}) &= \mathbb{E}_P[e^{c\tilde{f}}], \\ \tilde{f} &= f - \mathbb{E}_P[f]. \end{aligned} \tag{6}$$

The function $\log M_P(c; \tilde{f})$ stands for the cumulant-generating function (CGF) of the centralized observable $\tilde{f}$. The definition of the GO divergences involves the Donsker-Varadhan variational representation [23] of the KL, leading to bounds that are sharp and capture the worst-case model bias:

$$-\Xi_-(Q\|P; f) \leq \mathbb{E}_Q[f] - \mathbb{E}_P[f] \leq \Xi_+(Q\|P; f), \tag{7}$$

In addition, it was shown that the bounds in (7) remain sharp even when we are interested in ergodic observables:

$$f(x_{1:N}) = \frac{1}{N}\sum_{i=1}^{N} g(x_i),$$

for some function $g$ and data $\{x_i\}$; see [37] for details.

The theoretical properties of the GO divergences are inherited from its dependence on the CGF and the KL divergence. However, explicitly knowing the CGF can be a strong assumption when $P$ is not known, or is known up to a multiplicative constant, as is the case in Bayesian inference. Furthermore, estimation of the CGF via sampling is costly; as $c$ grows, we would need a large amount of data and/or computationally expensive multi-level Monte Carlo methods to meet a small tolerance for the variance of the estimator.

In Chapter 3, we take advantage of the properties of the model, $P$, and of the QoI, $f$, and relax the requirements of the GO bounds. To accomplish this, we turn to concentration inequalities [57, 13, 62] which make use of the concentration of measure phenomenon to provide bounds to the moment-generating function (MGF). Starting from the Hoeffding inequality [31], we show that we can derive divergences that retain the sharpness of the GO bounds (7) without requiring the computation of the CGF. In fact, concentration inequalities provide a systematic way to pick the information we want the bound to take into account about $f$ and $P$. As a result, we provide bias *guarantees* that use the available data and bound the bias over a whole class of QoIs. We demonstrate the usage of the bounds with a series of simple examples.

# CHAPTER 1

# INFORMATION METRICS FOR LONG-TIME ERRORS IN SPLITTING SCHEMES FOR STOCHASTIC DYNAMICS AND PARALLEL KINETIC MONTE CARLO

In this chapter[1], we propose an information-theoretic approach to analyze the long-time behavior of numerical splitting schemes for stochastic dynamics, focusing primarily on parallel kinetic Monte Carlo (PL-KMC) algorithms. Established methods for numerical operator splittings provide error estimates in finite-time regimes, in terms of the order of the local error and the associated commutator. Path-space information-theoretic tools such as the relative entropy rate allow us to control long-time error through commutator calculations. Furthermore, they give rise to an a posteriori representation of the error which can thus be tracked in the course of a simulation. Another outcome of our analysis is the derivation of a path-space information criterion for comparison (and possibly design) of numerical schemes, in analogy to classical information criteria for model selection and discrimination. In the context of parallel KMC, our analysis allows us to select schemes with improved numerical error and more efficient processor communication. We expect that such a path-space information perspective on numerical methods will be broadly applicable in stochastic dynamics, for both the finite and the long-time regime.

## 1.1 Background

Consider that the stochastic process of interest is an ergodic continuous time Markov Chain (CTMC) $X_t$ on a finite, but possibly still significantly large, state

[1]The contents of this chapter are published in the SIAM Journal of Scientific Computing [27] and appear here with permission.

space $S$. This stochastic process can be completely defined by its *transition rates*, $q(\sigma, \sigma')$, which describe the probability of an update from state $\sigma$ to state $\sigma'$ in an infinitesimal period of time. That is,

$$P(X_{t+\Delta t} = \sigma' | X_t = \sigma) = P_{\Delta t}(\sigma, \sigma') = q(\sigma, \sigma')\Delta t + o(\Delta t), \sigma \neq \sigma'. \tag{1.1}$$

Kinetic Monte Carlo (KMC) works by simulating the embedded Markov Chain $Y_n = X_{t_n}$, with jump times $t_n, t_n \sim \exp(\lambda)$. The parameter $\lambda(\sigma)$ is the total rate when the system is at state $\sigma$,

$$\lambda(\sigma) = \sum_{\substack{\sigma' \neq \sigma \\ \sigma' \in S}} q(\sigma, \sigma'). \tag{1.2}$$

This allows us to write the transition probabilities of the embedded Markov Chain $p(\sigma, \sigma') = q(\sigma, \sigma')/\lambda(\sigma)$. We can also define the infinitesimal generator $L$ that corresponds to the Markov chain as follows. First, consider $f$: bounded and continuous function on the state space $S$. Then, $L$ acts on $f$ at the state $\sigma$ as

$$L[f](\sigma) = \sum_{\sigma' \in S} q(\sigma, \sigma') \left( f(\sigma') - f(\sigma) \right). \tag{1.3}$$

Note that $L[\delta_{\sigma'}](\sigma) = q(\sigma, \sigma')$ for all states $\sigma, \sigma'$, where $\delta_{\sigma'}(\sigma) = \delta(\sigma, \sigma')$ is a Dirac probability measure. We shall also use the notation $L^k$ for the resulting operator after $k$ successive compositions of $L$. Because $L^k[\delta_{\sigma'}](\sigma) = L^{k-1}[L[\delta_{\sigma'}]](\sigma)$, we see that, for any $k$, $L^k[\delta_{\sigma'}](\sigma)$ is a computable object that depends on the transition rates.

Under fairly general conditions [39], the transition probability of the Markov process can be written as in semigroup form, i.e., $P_t(\sigma, \sigma') = e^{Lt}\delta_{\sigma'}(\sigma)$. In the case of interest to us, $L$ is going to be a bounded operator and such operators allow for a representation of the semigroup with a series expansion.

**Lemma 1.** *Let $L$ be a linear and bounded operator, $L : C_b(S) \to C_b(S)$, with $C_b(S)$ being the set of continuous and bounded functions on the space $S$. Then $L$ generates a uniformly continuous semigroup $e^{tL}$ which we can express in power series form.*

$$e^{tL} = \sum_{k=0}^{\infty} \frac{t^k}{k!} L^k. \tag{1.4}$$

*Proof.* This is a classical result for which many references exist; see, for example, Chapter 1, p. 2, of Pazy [54]. □

Thus, making use of Lemma 1, we can write the transition as

$$P_t(\sigma, \sigma') = e^{tL}\delta_{\sigma'}(\sigma) = \sum_{k=0}^{\infty} \frac{t^k}{k!} L^k[\delta'_\sigma](\sigma),\ \sigma, \sigma' \in S. \tag{1.5}$$

### 1.1.1 Constructing approximations by semigroup splitting

We will now give the foundations of approximations by splitting methods, as applied to the simulation of CTMCs, and proceed with how those ideas are applied in the case of PL-KMC.

As mentioned earlier, the transition probability of the Continuous Time Markov Chain (CTMC) of interest can be written as $e^{tL}\delta_{\sigma'}(\sigma)$. The goal is for us to design a splitting scheme that can approximate the action of $e^{tL}$. In our context, this leads to a new CTMC. One way to build such a scheme is to start with a splitting of the infinitesimal generator $L$ (1.3) into components $L_1, L_2$ with $L = L_1 + L_2$. Then, if we consider a positive $T$ and by using the Trotter product formula [60], we have

$$e^{TL} = \lim_{n\to\infty} (e^{T/nL_1} e^{T/nL_2})^n. \tag{1.6}$$

Correspondingly, if we now fix $n \in \mathbb{N}$ and set $\Delta t = T/n$, we can write approximations of $e^{TL}$ by using (1.6). For example, two such approximations are

$$
\begin{aligned}
e^{TL} &\simeq \left(e^{\Delta t L_1} e^{\Delta t L_2}\right)^n \text{ (Lie)}, \\
e^{TL} &\simeq \left(e^{\Delta t/2 L_1} e^{\Delta t L_2} e^{\Delta t/2 L_1}\right)^n \text{ (Strang)}.
\end{aligned}
\tag{1.7}
$$

Therefore for a one-step transition from $t = 0$ to $\Delta t$, (1.7) can be written as

$$
\begin{aligned}
e^{L\Delta t} &\simeq e^{\Delta t L_1} e^{\Delta t L_2}, \\
e^{L\Delta t} &\simeq e^{\Delta t/2 L_1} e^{\Delta t L_2} e^{\Delta t/2 L_1}.
\end{aligned}
\tag{1.8}
$$

Operator splittings can also be carried out with multiple components, such as $L = L_1 + L_2 + L_3 + L_4$. Such a splitting is used for two-dimensional (2D) lattice decompositions in SPPARKS [56]. All arguments can be simply extended to those cases, but we restrict to two components, $L_1, L_2$.

Throughout this work, we use $P_{\Delta t}(\sigma, \sigma')$ to denote the probability $e^{L\Delta t}\delta_{\sigma'}(\sigma)$ and $Q_{\Delta t}(\sigma, \sigma')$ for the approximations arising from splittings of the semigroup. Since $L$ is a bounded operator, we can express $P_{\Delta t}$ as expansion (1.5). If we pick $L_1, L_2$ so that they are also bounded, then we can express $Q_{\Delta t}$ as an expansion too. For example, for the Lie splitting

$$
\exp(\Delta t L_1)\exp(\Delta t L_2)\delta'_\sigma(\sigma) = \sum_{k=0}^{\infty} \frac{\Delta t^k}{k!} \left( k! \cdot \sum_{m=0}^{k} \frac{L_1^m}{m!} \cdot \frac{L_2^{k-m}}{(k-m)!} \right) \delta_{\sigma'}(\sigma), \tag{1.9}
$$

which we can show by multiplying the semigroup expansions of $\exp(\Delta t L_1)$ and $\exp(\Delta t L_2)$. Thus, if we use the notation

$$
L_Q^k := k! \cdot \sum_{m=0}^{k} \frac{L_1^m}{m!} \cdot \frac{L_2^{k-m}}{(k-m)!} \tag{1.10}
$$

we can write (1.9) in the form

$$
Q_{\Delta t}(\sigma, \sigma') = \sum_{k=0}^{\infty} \frac{\Delta t^k}{k!} L_Q^k[\delta_{\sigma'}](\sigma). \tag{1.11}
$$

By the definition of $L_Q^k$ in (1.10), $L_Q^0 = I$, $L_Q^1 = L$, $L_Q^2 = (L_1^2 + L_2^2 + 2L_1L_2)$, and so on, for the case of the Lie splitting. By a similar argument, we can write an expansion like (1.11) for other operator splitting approximations. In general, $L_Q$ is not a generator of a Markov process and, in that case, $L_Q^k$ is not equal $L_Q$ after $k$ compositions but is defined in the context of the expansion in (1.11). The slight abuse of notation allows us to compare the expansion of the exact process (1.5) with expansions of the approximating schemes of the form (1.11).

One way to compare the accuracy of using $Q_{\Delta t}$ as opposed to $P_{\Delta t}$ is to calculate the local error between expansion (1.5) and (1.11). As an example, here are the corresponding relations for the Lie and Strang splittings. We use $Q_{\Delta t}^{\text{Lie}}, Q_{\Delta t}^{\text{Strang}}$ for Lie and Strang, respectively. We will also use the notation $[L_1, L_2] := L_1L_2 - L_2L_1$ to denote the operator that captures the failure of $L_1$ and $L_2$ to commute. By using the expansions (1.5), (1.11), we can show that

$$P_{\Delta t}(\sigma, \sigma') = Q_{\Delta t}^{\text{Lie}}(\sigma, \sigma') + \frac{1}{2}[L_1, L_2]\delta_{\sigma'}(\sigma)\Delta t^2 + O(\Delta t^3), \tag{1.12}$$

$$P_{\Delta t}(\sigma, \sigma') = Q_{\Delta t}^{\text{Strang}}(\sigma, \sigma') + \frac{1}{24}\left([L_1, [L_1, L_2]] - 2[L_2, [L_2, L_1]]\right)\delta_{\sigma'}(\sigma)\Delta t^3 + O(\Delta t^4). \tag{1.13}$$

From relations (1.12) and (1.13), we observe that the Strang splitting has a better local error compared to Lie ($\Delta t^3$ versus $\Delta t^2$). Therefore, if we prescribe an error tolerance, the Strang scheme will be able to accommodate a larger $\Delta t$ than the Lie scheme. With a larger $\Delta t$, we will be able to take larger steps with the same tolerance during the simulation, and this is especially important for parallel KMC, as we strive for balance between error accumulation and efficiency.

To be able to discuss more general operator splitting approximations to $P_{\Delta t}$, we introduce the following helpful lemma.

**Lemma 2** (local order of error and commutator)**.** *Let* $P_{\Delta t}(\sigma, \sigma') = e^{L\Delta t}\delta_{\sigma'}(\sigma)$ *and* $Q_{\Delta t}(\sigma, \sigma')$ *an approximation of* $P_{\Delta t}$ *via a splitting scheme. Then, there is a function* $C : S \times S \to \mathbb{R}$ *and an integer* $p$*,* $p > 1$*, such that*

$$P_{\Delta t}(\sigma, \sigma') = Q_{\Delta t}(\sigma, \sigma') + C(\sigma, \sigma')\Delta t^p + o(\Delta t^p). \tag{1.14}$$

We will refer to $C(\sigma, \sigma') = (L^p - L_Q^p)\delta_{\sigma'}(\sigma)$ as the *commutator* and to $p$ as the *order of the local error.*

*Proof.* The result is immediate by using representations (1.5), (1.11), since for $\sigma, \sigma' \in S$,

$$P_{\Delta t}(\sigma, \sigma') - Q_{\Delta t}(\sigma, \sigma') = \sum_{k=0}^{\infty} \frac{\Delta t^k}{k!} \left(L^k - L_Q^k\right) [\delta'_\sigma](\sigma).$$

Then, $p$ is the smallest nonnegative integer such that $L^p \neq L_Q^p$. This of course implies that $L^k = L_Q^k$ for $k < p$. □

Equations (1.12) and (1.13) are examples of this lemma for the cases of the Lie and Strang splittings, respectively. Although in the case of Lie we were able to write the form of $L_Q^k$ explicitly for all $k$ (see Equation (1.10)), this is not a requirement and we only need to know $L_Q^p$ to compute the commutator and that object arises naturally when subtracting the two expansions, (1.5) and (1.11).

**Remark 3.** Relation (1.14) is central to the numerical analysis of splitting schemes, as it is the starting point to the derivation of upper bounds for the local and global error [5, 4, 30]. Even though our focus in this manuscript is on operator splitting schemes for parallel KMC, as long as an expression for the local error such as (1.14) exists, a similar analysis can be carried out for other types of schemes.

As we will see in the follow-up, the commutator has many desired properties. Since it is equal to $(L^p - L_Q^p)\delta_{\sigma'}[\sigma]$, and both $L^p[\delta_{\sigma'}](\sigma)$ and $L_Q^p[\delta_{\sigma'}](\sigma)$ depend on

the known transition rates $q$, the commutator is a *computable* object for every pair of states $(\sigma, \sigma')$. We will see in Section 1.4.1 that for parallel KMC the work required in order to compute the commutator can scale appropriately with the system size.

### 1.1.2 PL-KMC and splitting schemes

We consider the case of PL-KMC as an application of the ideas in the previous section concerning approximations by semigroup splitting. Further discussion on the ideas of this section can be found in Arampatzis et al. [5, 4].

Our main motivating example for PL-KMC is an interacting particle system. Let $\Lambda \subset \mathbb{Z}^d$ be a square lattice with $N$ sites. At each site of it, $x \in \Lambda$, we define an order parameter $\sigma(x) \in \Sigma = \{0, 1, \ldots, K\}$. This parameter can be, for example, the species that occupies the lattice site $x$. For instance, in the Ising model, $\sigma(x) = 0$ would imply that the lattice site $x$ is empty and $\sigma(x) = 1$ that a particle occupies $x$. The CTMC of interest is $\{\sigma_t\}_{t\geq 0}$, $\sigma_t = \{\sigma_t(x) : x \in \Lambda\}$, with state space $S = \Sigma^\Lambda$. At every $t$, $\sigma_t$ represents a snapshot of the different spins of the lattice. We can describe the dynamics of such a system by looking at the individual spin changes at different lattice sites. Two more properties that are common among such systems and which we will also assume is that the transitions between states of $\sigma_t$ are *localized* and that they only involve a finite number of lattice sites per transition step. Localization implies that the probability that a certain transition will happen (the order parameter of a finite collection of lattice sites will change) only depends on the values of $\sigma$ on a neighborhood around those lattice sites. In other words, transitions depend on local (neighborhood) rather than global (whole lattice) information (see Figure 1.1).

We can formalize localization by looking at the implication for the transition rates of the process $\sigma_t$. Following the notation introduced in [5], let us assume that at time $t$, $\sigma_t = \sigma$. Now, we can express the transition rate for a jump to a new state $\sigma^{x,\omega}$ as

$$q(\sigma, \sigma^{x,\omega}) = q(x, \omega; \sigma), \tag{1.15}$$

where $x \in \Lambda$ and $\omega$ is an index of the set of all possible configurations, $S_x$, that correspond to an update at a lattice neighborhood $\Omega_x$ of the site $x$. When the only allowed transition is spin-flipping, that is, starting with $\sigma$, we can only go to states $\sigma'$ that differ in the order parameter of one lattice site $x$, we will write $\sigma'$ as $\sigma^x$ to denote the resulting state after the transition. It follows that for $\sigma_t$ we have an infinitesimal generator:

$$L[f](\sigma) = \sum_{x \in \Lambda} \sum_{\omega \in S_x} q(x, \omega; \sigma) \left( f(\sigma^{x,\omega}) - f(\sigma) \right). \tag{1.16}$$

We can simulate the process $\sigma_t$ via standard KMC, as described in the beginning of Section 1.1. Then the system would progress in time steps $t_n \sim \exp(\lambda(\sigma))$, where $\lambda(\sigma)$ is the total rate when the system is at state $\sigma$, as defined in (1.2). Since the total rate scales with the size of the lattice and the magnitude of the transition rates, a large or highly reactive model would be simulated slowly by classical KMC. The

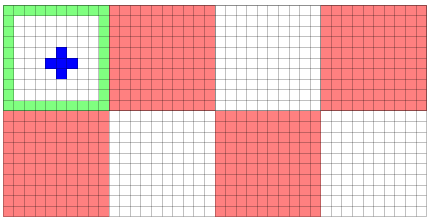

**Figure 1.1.** A checkerboard decomposition of a 2D lattice. Red sublattices correspond to group $G_1$ and white ones to $G_2$. For comparison, a nearest neighborhood region (n.n. region) is also shown (solid black cross). Transitions involving the center of that region only depend on the state of its nearest neighbors. So, if we pick the sublattices much larger than the size of an n.n. region, transitions in different sublattices belonging to the same group are independent. A site $x$ is said to belong to the boundary of its sublattice if part of its n.n. region is outside that sublattice (the green region is the collection of all such points for the first sublattice). If a transition occurs at such a site $x$, then an update needs to be made to the boundary information of all other sublattices for which $x$ belongs to an n.n. region.

goal then, as realized in [5], is for a fixed $\Delta t > 0$ to design an approximation to the exact process $e^{\Delta t L}$ via a splitting method in such a way that allows for asynchronous computations.

To begin, we note that any decomposition of the lattice into nonoverlapping sublattices $\Lambda_i$ also induces a decomposition of the generator (1.16), that is,

$$L[f](\sigma) = \sum_{i=1}^{n} \sum_{x \in \Lambda_i} \sum_{\omega \in S_x} q(x, \omega; \sigma) \left( f(\sigma^{x,\omega}) - f(\sigma) \right). \tag{1.17}$$

Due to the localization of the system, we can decompose the lattice $\Lambda$ into $n$ sublattices, $\Lambda_i$, so that transitions in some sublattices are independent from transitions in others; see Figure 1.1. With two groups, $G_1 = \{\Lambda_i : i \text{ even}\}, G_2 = \{\Lambda_i : i \text{ odd}\}$, we can split $L$ into

$$\begin{aligned} L_j[f](\sigma) &:= \sum_{x \in G_j} \sum_{\omega \in S_x} q(x, \omega; \sigma) \left( f(\sigma^{x,\omega}) - f(\sigma) \right), \ j = 1, 2, \\ L[f](\sigma) &= L_1[f](\sigma) + L_2[f](\sigma). \end{aligned} \tag{1.18}$$

Thus, by the formulas in (1.18), we can use the ideas of the previous section to construct splitting approximations to $e^{L\Delta t}$. Those can also be interpreted as computation schedules for the parallel algorithm. Such schedules set two attributes of the simulation: (a) in what order to simulate the two groups asynchronously and (b) for how much time to simulate each group per time step (which the user controls with the $\Delta t$ parameter). A demonstration of how PL-KMC works is shown in Figure 1.2.

In general, the larger the $\Delta t$, the less different processes need to communicate to resolve inconsistencies during a run. This is a fact for any simulation algorithm that can be expressed in the above operation splitting framework, e.g., SPPARKS and others [5]. Since communication is the usual bottleneck of PL-KMC algorithms, a practitioner would like to pick $\Delta t$ as large as possible, given a fixed tolerance. One of the important insights of the analysis in [4] is that the commutator controls this

relationship. Simply put, a small $C(\cdot,\cdot)$ (as defined in Lemma 2) allows for a larger step size $\Delta t$.

## 1.2 Information metrics for comparing dynamics at long times

We will now introduce the main tools from information theory. In later sections, our focus will be to compare the exact process, $X_t$, and an approximation of it, $Y_t$, via their $\Delta t$-skeleton subprocesses. That is, given a fixed $\Delta t > 0$ and $M \in \mathbb{N}$, we look at the discrete-time Markov processes $X_{n\Delta t}$ and $Y_{n\Delta t}$ for $n \in \{0, \ldots, M\Delta t\}$ and $T = M\Delta t$. For this reason, we now introduce the information-theoretical tools we will use for discrete-time processes.

Consider two discrete-time Markov processes $X_n$ and $Y_n$ on a countable state space $S$ with transition probabilities $P$ and $Q$, respectively. We also assume that for each process there exists a corresponding unique stationary distribution $\mu_P$ and $\mu_{\Delta t}$. Assuming $X_0$ ($Y_0$) is distributed according to $\mu_P$ ($\mu_{\Delta t}$), we can then calculate the probability of a specific path for each process. For example, if we fix a positive

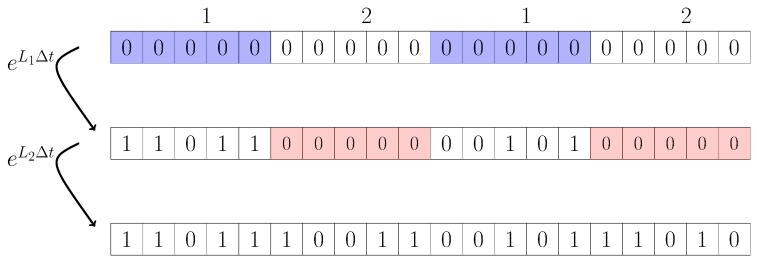


**Figure 1.2.** One step of PL-KMC in the $1D$ case, where all of the spin values are set to zero initially while using the Lie splitting. After the lattice is decomposed into nonoverlapping sublattices, here blue (indexed as 1) and red (indexed as 2), the algorithm proceeds by first simulating all blue sublattices independently by standard KMC until a time $t = \Delta t$ is reached for all of them. Once that is done, the lattices in the second group are simulated in the same way. This results to the process $\sigma_t$ on the whole lattice being propagated forward in time by $\Delta t$. Between the simulation of each group, communication between the processes is required in order to correct for the mismatch on the boundaries of the sublattices. The resulting error due to the mismatch is controlled by the commutator $C$ [4].

integer $M$, let $T = M\Delta t$ and pick an $\vec{x} \in S^M$, then we have

$$P_{0:T}(\vec{x}) = P(X_T = x_M, \ldots, X_0 = x_0) = \mu_P(x_0)P(x_0, x_1)\cdots P(x_{M-1}, x_M),$$

i.e., the path probability for $P$ can be factorized according to the Markov property. Similarly, by changing $P$ to $Q$, we can calculate the path probability for $Y_n$.

If we used a path of length $T$ from the process $Y_n$ to compute quantities of interest (QOIs), e.g., expected values on the path, how much information would be lost compared to using a path of length $T$ from $X_n$. This is a central question in coding theory and one way to quantify the information loss is through the idea of relative entropy,

$$R(Q_{0:T}|P_{0:T}) := \sum_{\vec{x}\in S^M} Q_{0:T}(\vec{x}) \log \frac{Q_{0:T}(\vec{x})}{P_{0:T}(\vec{x})}. \tag{1.19}$$

Our definition here is with respect to the path measures $P_{0:T}, Q_{0:T}$, but we can apply the relative entropy to more general probability measures too. For this object to be properly defined, we need to have that $Q_{0:T}$ is absolutely continuous with respect to $P_{0:T}$, that is, $P_{0:T}(\vec{x}) = 0$ implies $Q_{0:T}(\vec{x}) = 0$. Other important properties of (1.19) are the following: (1). $R(Q_{0:T}|P_{0:T}) \geq 0$ for any $Q_{0:T}, P_{0:T}$ (Gibbs' inequality); (2) $R(Q_{0:T}|P_{0:T}) = 0 \Leftrightarrow P_{0:T} = Q_{0:T}$. Note though that the relative entropy does not qualify as a metric in the classical sense, as it is not symmetric and does not satisfy the triangle inequality. It can, however, still be thought of as a notion of a distance between distributions and is useful as a building block for other information measures. For a more complete exposition on relative entropy and its properties, see [18].

Although the pathwise relative entropy is a suitable quantity to measure the similarity of the two path measures, it is computationally demanding to calculate, especially in the case of parallel KMC, where we do not have $Q_{0:T}$ and $P_{0:T}$ explicitly. For this reason, we look at a related object, the relative entropy per unit time, or

relative entropy rate (RER). Given a probability measure $\nu_0$, $\nu_0(\vec{x}) = \nu_0(x_0), \vec{x} \in S^T$, the RER with respect to $\nu_0$ is defined as

$$H_{\nu_0}(Q|P) := \sum_{\vec{x} \in S^M} \nu_0(\vec{x}) Q(x_0, x_1) \log \frac{Q(x_0, x_1)}{P(x_0, x_1)}. \tag{1.20}$$

Given another measure $\mu_0$, we can use the chain rule for the relative entropy [18] to relate relative entropy and RER as

$$\begin{aligned} R(Q_{0:T}|P_{0:T}) &= R(\mu_0|\nu_0) + \sum_{i=1}^{M} H_{\nu_i}(Q|P), \\ \nu_k(x_0, \ldots, x_{k-1}) &= \nu_0(x_0) \prod_{m=1}^{k-1} Q(x_{m-1}, x_m). \end{aligned} \tag{1.21}$$

In particular, when sampling from the stationary distribution corresponding to $Q$, that is, $\nu_0 = \mu_{\Delta t}$, then $H_{\nu_i} = H_{\mu_{\Delta t}} = H$ for all $i$. Then,

$$H(Q|P) = \sum_{x_0, x_1 \in S} \mu_{\Delta t}(x_0) Q(x_0, x_1) \log \frac{Q(x_0, x_1)}{P(x_0, x_1)}. \tag{1.22}$$

This also simplifies (1.21) to

$$R(Q_{0:T}|P_{0:T}) = M \cdot H(Q|P) + R(\mu_{\Delta t}|\mu_P). \tag{1.23}$$

In (1.23), $R(\mu_{\Delta t}|\mu_P)$ is the relative entropy of $\mu_{\Delta t}$ with respect to $\mu_P$, capturing the loss of information between the exact and approximate stationary distribution. Note that $R(\mu_{\Delta t}|\mu_P)$ does not depend on the length of the path. Instead, the term that quantifies the dependence on $T$ is $H(Q|P)$. Therefore, any difference between the two stationary measures becomes negligible for large times, which is a first advantage to studying the pathwise relative entropy through the simpler RER.

### 1.2.1 Information metrics and observables

Further justification for the fact that the RER is the right quantity to track can be given by considering time-averaged observables. For instance, if $f$ is a function of the state space, then such an observable would be

$$M \cdot F_M(\{X_n : n = 0, \dots, M-1\}) = \sum_{k=0}^{M-1} f(X_k).$$

An important performance metric for the approximation is the weak error:

$$|\mathbb{E}_{P[0,T]}[F_M] - \mathbb{E}_{Q[0,T]}[F_M]|, \text{ where } T = M\Delta t. \tag{1.24}$$

In recent work [22], uncertainty quantification bounds have been developed for the weak error that are of the form

$$\begin{aligned} \Xi_-(Q_{[0,T]} \| P_{[0,T]}; M \cdot F_M)/M \leq \mathbb{E}_{P[0,T]}[F_M] - \mathbb{E}_{Q[0,T]}[F_M] \\ \leq \Xi_+(Q_{[0,T]} \| P_{[0,T]}; M \cdot F_M)/M. \end{aligned} \tag{1.25}$$

The quantities $\Xi_\pm(Q_{[0,T]} \| P_{[0,T]}; M \cdot F_M)$ are defined as goal-oriented divergences [22], taking into account the observable $F$, and such that $\Xi_\pm(Q_{[0,T]} \| P_{[0,T]}; M \cdot F_M) = 0$, if $Q_{[0,T]}, = P_{[0,T]}$ or $f$ is deterministic. Note that the bound in (1.25) is robust in the following sense: if we consider a positive $\eta$ and all $Q_{\Delta t}$ such that $R(Q_{\Delta t} | P_{\Delta t}) < \eta$, then the upper bound in (1.25) is attained; see Theorem 3.4 in [17], as well as [36].

Dividing (1.25) by $M$ and letting $M$ go to infinity gives an inequality with respect to the stationary measures $\mu_{\Delta t}, \mu_P$ of the scheme, $Q_{\Delta t}$, and the exact process, $P_{\Delta t}$, respectively:

$$\xi_-(Q_{\Delta t} \| P_{\Delta t}; f) \leq \mathbb{E}_{\mu_{\Delta t}}[f] - \mathbb{E}_{\mu_P}[f] \leq \xi_+(Q_{\Delta t} \| P_{\Delta t}; f), \tag{1.26}$$

where $\xi_\pm(Q_{\Delta t}\|P_{\Delta t};f) = \lim_{M\to\infty}\Xi_\pm(Q_{0:T}\|P_{0:T};F)/M$. But $\xi_\pm$ also admit a variational representation as

$$\begin{aligned}\xi_+(Q_{\Delta t}\|P_{\Delta t};f) &= \inf_{c\geq 0}\left\{\frac{1}{c}[\lambda_{Q_{\Delta t},P_{\Delta t}}(c)+H(Q_{\Delta t}\|P_{\Delta t})]\right\},\\ \xi_-(Q_{\Delta t}\|P_{\Delta t};f) &= \sup_{c\geq 0}\left\{-\frac{1}{c}[\lambda_{Q_{\Delta t},P_{\Delta t}}(-c)+H(Q_{\Delta t}\|P_{\Delta t})]\right\},\end{aligned} \tag{1.27}$$

with $\lambda_{Q_{\Delta t},P_{\Delta t}}(c)$ in (1.27) to be the logarithm of the maximum eigenvalue of the matrix with entries $P_{\Delta t}(x,y)\exp(c\cdot(f(y)-\mathbb{E}_{\mu_P}[f]))$ (see [36] for details). Especially when $H(Q_{\Delta t}|P_{\Delta t})$ is small and through the asymptotic expansion of $\xi_\pm$, an upper bound for the weak error at stationarity can be given (following the ideas in [22, 36]):

$$|\mathbb{E}_{\mu_{\Delta t}}[f]-\mathbb{E}_{\mu_P}[f]| \leq \sqrt{\upsilon_{\mu_P}(f)}\sqrt{2H(Q_{\Delta t}|P_{\Delta t})}+O(H(Q_{\Delta t}|P_{\Delta t})), \tag{1.28}$$

$$\upsilon_{\mu_P}(f) = \sum_{k=-\infty}^{\infty}\mathbb{E}_{\mu_P}[f(X_k)f(X_0)]. \tag{1.29}$$

Inequality (1.28) connects the long-time loss of accuracy that the weak error captures with the RER and $\upsilon_{\mu_P}(f)$, which is the integrated auto-correlation function for the observable $f$ and a quantity we can estimate during the simulation. As a consequence of (1.28), any further results on the asymptotic behavior of $H(Q_{\Delta t}|P_{\Delta t})$ with respect to $\Delta t$ can be simply translated to the weak error point of view.

## 1.3 Long-time error behavior of splitting schemes

In this section, we compare the RER between two different processes. One of them will always be the $\Delta t$-skeleton process derived from the CTMC we wish to simulate, with transition probability

$$P_{\Delta t}(\sigma,\sigma') = e^{L\Delta t}\delta_{\sigma'}(\sigma). \tag{1.30}$$

This exact $\Delta t$-process will be compared with the $\Delta t$-skeleton process derived from an operator splitting of (1.30). Such approximations will be denoted with $Q_{\Delta t}$. We note here that the discretization (1.30) of the original Markov process with semigroup $e^{tL}$ with respect to $\Delta t$ is carried out only as a means to compare the original process with the approximations $Q_{\Delta t}$. The transition kernel $P_{\Delta t}$ is just a particular instance of the transition matrix of the continuous Markov process with semigroup $P_t = e^{tL}$, so there is no approximation error in (1.30). In fact, using the $\Delta t$-skeleton corresponds to subsampling from the CTMC at every $\Delta t$.

Our goal is to show the dependence of the RER on various quantities of interest that are usually computed for short-time error analysis. We will see that the commutator, the order of the local error, and other quantities make an appearance in the asymptotic results we develop. We limit our discussion to the case that $\Delta t$ is in $(0, 1]$, as this is the interval where splitting schemes are most accurate and so fair comparisons can be made. We also assume throughout this section that $L$ is a bounded operator. We will often refer to the splittings previously discussed, Lie and Strang, which define discrete processes with transition probabilities

$$\begin{aligned} Q_{\Delta t}^{\text{Lie}}(\sigma, \sigma') &= e^{L_1 \Delta t} e^{L_2 \Delta t} \delta_{\sigma'}(\sigma), \\ Q_{\Delta t}^{\text{Strang}}(\sigma, \sigma') &= e^{L_1 \Delta t/2} e^{L_2 \Delta t} e^{L_1 \Delta t/2} \delta_{\sigma'}(\sigma). \end{aligned} \tag{1.31}$$

Here $L$ is the original generator and $L = L_1 + L_2$ with $L_1, L_2$ assumed bounded as operators. For instance, in the case of parallel KMC, $L_1, L_2$ will be imposed by the domain decomposition of the lattice; see Figure 1.1.

Before we move on to the analysis, we need to address a last issue. Recall that our main tool will be asymptotic expansions of the RER with respect to $\Delta t$. We will then use those to do comparisons for different $\Delta t$, so it is important to first account for the scaling of RER with respect to that parameter. The situation can be best

illustrated by the worst-case scenario, when the order of the local error between two Markov semigroups, $Q^A_{\Delta t},\ Q^B_{\Delta t}$, is equal to one.

**Lemma 4.** *Let $L_A, L_B$ be bounded generators of Markov processes, $L_A \neq L_B$, with corresponding transition probabilities $Q^A_{\Delta t}, Q^B_{\Delta t}$ . Then,*

$$H(Q^B_{\Delta t}|Q^A_{\Delta t}) = O(\Delta t).$$

*Proof.* The proof follows the ideas in Theorem 7. The argument is provided in Appendix A.2.2. □

**Remark 5.** Using Lemma 4, we can readily see that given an operator splitting scheme $Q_{\Delta t}$ that approximates the exact $P_{\Delta t}$, we expect a scaling at least of the type $H(Q_{\Delta t}|P_{\Delta t}) = O(\Delta t)$. To correct for the $\Delta t$ scaling, we will instead work with a $\Delta t$-normalized RER. That is, we redefine the RER as

$$H(Q_{\Delta t}|P_{\Delta t}) := \frac{1}{\Delta t} \sum_{\sigma,\sigma'} \mu_{\Delta t}(\sigma) Q_{\Delta t}(\sigma,\sigma') \log\left(\frac{Q_{\Delta t}(\sigma,\sigma')}{P_{\Delta t}(\sigma,\sigma')}\right). \tag{1.32}$$

We wish to use the RER (see (1.32)) to study the long-time loss of information between $Q_{\Delta t}$ and $P_{\Delta t}$. However, in the case of parallel KMC, those are difficult to calculate explicitly, hence we turn to asymptotic expansions instead. We will see that the terms in those expansions depend on the transition rates and, under suitable ergodic assumptions, can be estimated during the simulation.

## 1.4 RER analysis for parallel KMC

We will now study an example from a class of interacting particle systems, limiting our discussion to the Lie and Strang splittings. Given two states $\sigma, \sigma' \in S$ and $x$ lattice site, $\sigma(x) \in \{0, 1\}$, we have that the transition rates $q$ are

$$q(\sigma,\sigma') = \begin{cases} q(\sigma,\sigma^x) > 0, & \sigma' = \sigma^x, \\ 0, & \text{otherwise.} \end{cases} \tag{1.33}$$

The rates in (1.33) provide a particular example of an adsorption/desorption system. Other mechanisms can be incorporated into (1.33), such as diffusion or reactions with multiple components or with particles that have many degrees of freedom [5].

Given a lattice $\Lambda$ with $N$ sites, we are interested in simulating the process $\sigma_t = \{\sigma_t(x) : x \in \Lambda\}$ in parallel with an operator splitting method, so we apply the ideas in Section 1.1.2 to that end. We first decompose the lattice into nonoverlapping sublattices (see Figure 1.1) and this induces a decomposition of the generator into new generators $L_1, L_2$ as in (1.18). Then, for any $T > 0$, the adsorption/desorption system can be simulated in $[0, T]$ using the parallel KMC algorithm. From the short-time error analysis, we can control the error by computing the commutator, $C(\cdot,\cdot)$, and the order of the local error that corresponds to the operator splitting scheme we use. For example, we know that for the Lie splitting that order is $p = 2$ and $C(\sigma,\sigma') = [L_1, L_2]\delta_{\sigma'}(\sigma)/2$ (see Lemma 2 and (1.12)). By using the properties of the generators $L_1, L_2$ along with our assumption in (1.33), we can show that

$$\begin{aligned} C(\sigma,\sigma') = [L_1, L_2]\delta_{\sigma'}(\sigma)/2 = \frac{1}{2}\sum_{x,y\in\Lambda} & f_1(x,y;\sigma)\delta_{\sigma'}(\sigma^{x,y}) - f_2(x,y;\sigma)\delta_{\sigma'}(\sigma^x) \\ & - \frac{1}{2}\sum_{x,y\in\Lambda} f_3(x,y;\sigma)\delta_{\sigma'}(\sigma^y), \end{aligned} \tag{1.34}$$

where $f_1, f_2$, and $f_3$ only depend on the transition rates $q$. We recall here that $\sigma^{x,y}$ stands for the resulting state $\sigma'$ after a spin-flip of an initial state $\sigma$ at lattice sites $x, y, x \neq y$. A full description of the above formula along with a proof can be found in the appendix.

**Remark 6.** Formula (1.34) for the Lie commutator has two important properties. First, it is computable for any pair $(\sigma,\sigma') \in S \times S$ as it only depends on the transition

rates $q$. Second, it is surely equal to zero if $\sigma' \neq \sigma^{x,y}$ and $\sigma' \neq \sigma^x$ for all $x, y \in \Lambda, x \neq y$, due to the $\delta_{\sigma'}$ appearing in the different sums. We will also see that the sum in (1.34) needs to be evaluated only for the neighboring lattice sites $x, y$ that are not both in the same group. For instance, in Figure 1.1, we would only need to evaluate the sum over the green boundary regions of every sublattice, which makes the computation of the commutator much simpler (see Remark 8 for a complexity analysis). Those properties hold for commutators of other operator splitting schemes too; see [4] and Section 1.7.

To study the asymptotic behavior of the RER, we will need to quantify the dependence of various combinations of $P_{\Delta t}$ and $Q_{\Delta t}$ to $\Delta t$. To this end, we use the following facts, both of which stem from Lemma 2:

$$P_{\Delta t}(\sigma, \sigma') - Q_{\Delta t}(\sigma, \sigma') = C(\sigma, \sigma')\Delta t^p + o(\Delta t^p), \tag{1.35}$$

$$P_{\Delta t}(\sigma, \sigma') + Q_{\Delta t}(\sigma, \sigma') = 2\delta_{\sigma'}(\sigma) + 2q(\sigma, \sigma')\Delta t + o(\Delta t) \tag{1.36}$$

$$= 2Q_{\Delta t}(\sigma, \sigma') + C(\sigma, \sigma')\Delta t^p + o(\Delta t^p). \tag{1.37}$$

We are now able to write an asymptotic result for RER for the Lie and Strang operator splittings in parallel KMC under the assumption in relation (1.33).

**Theorem 7.** *Let $\Delta t \in (0, 1)$ and $\sigma_{n\Delta t}$ on the lattice $\Lambda$ with transition probability $P_{\Delta t}(\sigma, \sigma') = e^{L\Delta t}\delta_{\sigma'}(\sigma)$ for $\sigma, \sigma' \in S$. Then, let $L_1 + L_2$ be a splitting of $L$ based on a decomposition of the lattice $\Lambda$. Assuming that property* (1.33) *holds for the rates, if there exists a state $\sigma \in S$ and lattice sites distinct $x, y$ such that the Lie commutator $C(\sigma, \sigma^{x,y}) \neq 0$, we have that*

$$H(Q^{\text{Lie}}_{\Delta t} | P_{\Delta t}) = O(\Delta t^1) \text{ (Lie).} \tag{1.38}$$

*Similarly, if there exists a state $\sigma \in S$ and distinct lattice sites $x, y, z$ such that $C(\sigma, \sigma^{x,y,z}) \neq 0$,*

$$H(Q_{\Delta t}^{\text{Strang}} | P_{\Delta t}) = O(\Delta t^2) \text{ (Strang).} \tag{1.39}$$

*Proof.* We will first show the result for the Lie case and then note the differences in the proof for the Strang case. Thus, we denote $Q_{\Delta t}^{\text{Lie}}$ by $Q_{\Delta t}$ and $\mu_{\text{Lie}}$ by $\mu_{\Delta t}$ and consider a $\Delta t \in (0, 1)$. As we wish to construct an asymptotic expansion for the RER in (1.32), we first need to expand the logarithm. Given a positive $x$ and by the definition of $\tanh^{-1}$,

$$\log(x) = 2\text{atanh}\left(\frac{x-1}{x+1}\right) = 2\sum_{k=0}^{\infty} \frac{1}{2k+1}\left(\frac{x-1}{x+1}\right)^{2k+1}. \tag{1.40}$$

This expansion of the logarithm converges for every $x > 0$, as can be seen by applying the root convergence test. Thus, expanding the logarithm part of the RER, we get

$$\begin{aligned} \Delta t \cdot H(Q_{\Delta t} | P_{\Delta t}) = &- 2\sum_{\sigma,\sigma'} \mu_Q(\sigma) Q_{\Delta t}(\sigma, \sigma') \frac{P_{\Delta t}(\sigma, \sigma') - Q_{\Delta t}(\sigma, \sigma')}{Q_{\Delta t}(\sigma, \sigma') + P_{\Delta t}(\sigma, \sigma')} \\ &+ 2\sum_{\sigma,\sigma'} \mu_Q(\sigma) J(\Delta t; \sigma, \sigma'), \end{aligned} \tag{1.41}$$

$$J(\Delta t; \sigma, \sigma') := Q_{\Delta t}(\sigma, \sigma') \sum_{k=1}^{\infty} \frac{1}{2k+1}\left(\frac{Q_{\Delta t}(\sigma, \sigma') - P_{\Delta t}(\sigma, \sigma')}{Q_{\Delta t}(\sigma, \sigma') + P_{\Delta t}(\sigma, \sigma')}\right)^{2k+1}. \tag{1.42}$$

We will study the asymptotic behavior of both parts of the RER in (1.41). First, applying (1.36) to the denominator of the fraction in (1.41) and carrying out the simplifications, we have

$$\begin{aligned} \Delta t \cdot H(Q_{\Delta t} | P_{\Delta t}) = &- 2\sum_{\sigma,\sigma'} \mu_{\Delta t}(\sigma) \left(P_{\Delta t}(\sigma, \sigma') - Q_{\Delta t}(\sigma, \sigma') + G(\Delta t; \sigma, \sigma')\right) \\ &+ 2\sum_{\sigma,\sigma'} \mu_{\Delta t}(\sigma) J(\Delta t; \sigma, \sigma'). \end{aligned} \tag{1.43}$$

Now, since $Q_{\Delta t}, P_{\Delta t}$ are transition probabilities, $\sum_{\sigma' \in S} P_{\Delta t}(\sigma, \sigma') - Q_{\Delta t}(\sigma, \sigma') = 0$ for all $\sigma \in S$, and thus the corresponding part of (1.43) is zero. To progress, we need to study the dependence on $\Delta t$ of $J, G$. First, for $G$ in (1.43),

$$G(\Delta t; \sigma, \sigma') = \frac{(P_{\Delta t}(\sigma, \sigma') - Q_{\Delta t}(\sigma, \sigma'))C(\sigma, \sigma')\Delta t^2}{(2Q_{\Delta t}(\sigma, \sigma') + \Delta t^2 C(\sigma, \sigma') + o(\Delta t^2))} + o(\Delta t^2). \tag{1.44}$$

To expose the dependence of the numerator of (1.44) to $\Delta t$, we use (1.35) to get

$$G(\Delta t; \sigma, \sigma') = \frac{(C(\sigma, \sigma'))^2}{2Q_{\Delta t}(\sigma, \sigma') + \Delta t^2 C(\sigma, \sigma') + o(\Delta t^2)} \Delta t^4 + o(\Delta t^2). \tag{1.45}$$

We wish to show that $G(\Delta t; \sigma, \sigma') = O(\Delta t^2)$. From the explicit form of the commutator in (1.34) and Remark 6, we can see that we need to study $G$ only in the cases that $\sigma' = \sigma^x$ or $\sigma' = \sigma^{x,y}$, given a state $\sigma$ and lattice sites $x, y$, since otherwise $C(\sigma, \sigma') = 0$. Let us consider $\sigma' = \sigma^{x,y}$. Since the order of the local error is equal to two, from expansion (1.11) and the fact that $L_Q[\delta_{\sigma^{x,y}}](\sigma) = L[\delta_{\sigma^{x,y}}](\sigma)$ and $L[\delta_{\sigma^{x,y}}] = q(\sigma, \sigma^{x,y}) = 0$ (see the property in (1.33)), we have

$$Q_{\Delta t}(\sigma, \sigma^{x,y}) = \frac{\Delta t^2}{2} L_Q^2[\delta_{\sigma'}](\sigma) + o(\Delta t^2). \tag{1.46}$$

Thus, applying (1.46) to the denominator of (1.45),

$$\begin{aligned} G(\Delta t; \sigma, \sigma^{x,y}) &= \frac{(C(\sigma, \sigma^{x,y}))^2}{\Delta t^2 \cdot (L_Q^2[\delta_{\sigma^{x,y}}](\sigma) + C(\sigma, \sigma^{x,y})) + o(\Delta t^2)} \Delta t^4 + o(\Delta t^2) \\ &= \frac{(C(\sigma, \sigma^{x,y}))^2}{L_Q^2[\delta_{\sigma^{x,y}}](\sigma) + C(\sigma, \sigma^{x,y})} \Delta t^2 + o(\Delta t^2). \end{aligned} \tag{1.47}$$

By similar calculations, we can show that $G(\sigma, \sigma^x) = O(\Delta t^3)$, if $C(\sigma, \sigma^x) \neq 0$ for that $x \in \Lambda$. Regardless, this would be a lower order, since $\Delta t < 1$. Thus, $G(\Delta t; \sigma, \sigma')$ is indeed of order $\Delta t^2$. Next, we will account for $J(\Delta t; \sigma, \sigma')$. If $\sigma' = \sigma^{x,y}$, then

$$J(\Delta t; \sigma, \sigma^{x,y}) = Q_{\Delta t}(\sigma, \sigma^{x,y}) \sum_{k=1}^{\infty} \frac{1}{2k+1} \left( \frac{Q_{\Delta t}(\sigma, \sigma^{x,y}) - P_{\Delta t}(\sigma, \sigma^{x,y})}{Q_{\Delta t}(\sigma, \sigma^{x,y}) + P_{\Delta t}(\sigma, \sigma^{x,y})} \right)^{2k+1}. \tag{1.48}$$

Because $Q_{\Delta t}(\sigma, \sigma^{x,y}) = O(\Delta t^2)$ and $Q_{\Delta t}(\sigma, \sigma^{x,y}) \pm P_{\Delta t}(\sigma, \sigma^{x,y}) = O(\Delta t^2)$, we get

$$J(\Delta t; \sigma, \sigma^{x,y}) = O(\Delta t^2),$$

since, for $\sigma' = \sigma^x$, $J(\Delta t; \sigma, \sigma^x) = O(\Delta t^4)$ and this is a lower order when $\Delta t < 1$. Therefore, $H(Q_{\Delta t}|P_{\Delta t}) = O(\Delta t^1)$. Note that all of the terms of the series in (1.48) contribute a term of order $\Delta t^2$, so the coefficient of $\Delta t^2$ in the asymptotic expansion of the RER will be a result of the summation of all those terms.

Finally, we discuss the differences in our argument for the proof of the Strang case. First, the order of the local error for Strang is $p = 3$, so every time we use formula (1.35) in the proof, we would introduce a term of order $\Delta t^3$ instead of $\Delta t^2$. Then, using an expression for $C(\cdot, \cdot)$ similar to (1.34) but for the Strang case, we would show that

$$J(\Delta t; \sigma, \sigma^{x,y,z}) = O(\Delta t^3) = G(\Delta t; \sigma, \sigma^{x,y,z})$$

for $x, y, z \in \Lambda$ and $x \neq y \neq z$. This would then give the result for Strang. □

### 1.4.1 Building biased a posteriori estimators for the RER

Theorem 7 shows that the long-time accuracy with respect to the RER of the two operator spllitting schemes, Lie and Strang, scales with $\Delta t$ in the same way the global error does. However, it also exposes the first terms in the asymptotic expansion of the RER for Lie and Strang. Essentially,

$$H(Q^{\text{Lie}}_{\Delta t}|P_{\Delta t}) = A\Delta t + o(\Delta t), \tag{1.49}$$

$$H(Q^{\text{Strang}}_{\Delta t}|P_{\Delta t}) = B\Delta t^2 + o(\Delta t^2), \tag{1.50}$$

where $A, B$ are the corresponding highest-order RER coefficients. Those have an explicit form that depends on the system one wishes to simulate and the commuta-

tor $C(\sigma, \sigma')$ corresponding to the scheme. We focus on the case of the Lie operator splitting, though similar comments can also be made for Strang. For systems with transition rates satisfying the property in (1.33), the highest-order coefficient $A$ appearing in (1.49) has the form

$$A = \sum_{\sigma} \mu_{\text{Lie}}(\sigma) \sum_{x,y \in \Lambda} C_{\text{Lie}}(\sigma, \sigma^{x,y}) F_{\text{Lie}}(\sigma, \sigma^{x,y}), \tag{1.51}$$

where $C_{\text{Lie}}$ is the Lie commutator (see (1.12)) and $F_{\text{Lie}}$ is a quantity that depends on the splitting (see (A.1) and (A.3) in the appendix for examples on how this $F$ can look for different splittings). Both $C$ and $F$ can be expressed in terms of the transition rates of the process $q$, i.e., they are computable for any state $\sigma$ and $x, y \in \Lambda$. Therefore, $A$ in (1.51) can be estimated via an ergodic average when simulating with the Lie scheme and hence, for small $\Delta t$, $H(Q^{\text{Lie}}_{\Delta t} | P_{\Delta t}) \simeq A \Delta t$.

At first glance, computing coefficient (1.51) involves work that scales with the size of the lattice. However, it was shown in Lemma 5.15 of [4] that the commutator only depends on the boundary regions between sublattices (see Figure 1.1). We will continue this discussion in Section 1.5, where we consider an adsorption-desorption system. We will also see that, apart from a comparison of the schemes in terms of the long-time loss of information, the estimators of RER can also be of use in tuning parameters of the scheme ($\Delta t$, domain decomposition, etc.). We will then consider the behavior of the RER when simulating other systems in Section 1.7.

## 1.5 Error versus communication and time-step selection

In this section, we explore the balance between numerical error and processor communication in parallel KMC, in the context of a specific example. Let us assume a bounded 2D lattice, $\Lambda \subset \mathbb{Z}^2$ with $100 \times 100$ sites. At each site $x$, we have a spin variable, $\sigma(x) \in \Sigma = \{0, 1\}$, with $\sigma(x) = 0$ denoting an empty site and $\sigma(x) = 1$

an occupied one. Our model in this case is going to be an *adsorption-desorption* one, although the analysis would similarly apply for other mechanisms (diffusions, reactions, etc.; see [5] for more details). The transition rates we will use correspond to spin-flip Arrhenius dynamics. Given a lattice site $x$, we may also define the nearest-neighbor set $\Omega_x = \{z \in \Lambda : |z - x| = 1\}$. The transitions rates are then

$$q(\sigma, \sigma^x) = q(x, \sigma) = c_1(1 - \sigma(x)) + c_2\sigma(x)e^{-\beta U(x)}, \tag{1.52}$$

$$U(x) = J_0 \sum_{y \in \Omega_x} \sigma(y) + h, \tag{1.53}$$

where $c_1, c_2, -\beta, J_0$, and $h$ are constants that can be tuned to generate different dynamics. We recall that $\sigma^x$ denotes the result of a spin-flip at lattice position $x$ if we start from state $\sigma$. Note that the transition rates (1.52) have the property (1.33). When considering a jump from $\sigma$ to $\sigma^x$, $q$ only depends in the spin values of the sites close to $x$ (through $U(x)$). Since transitions are localized, we can thus employ a geometrical decomposition of the lattice, as described in Section 1.1.1, and simulate the system in parallel. To accomplish this, we used Sandia Labs' SPPARKS code, a kinetic Monte Carlo simulator [56].

From Table 1.1 and Remark 8, we can see that the cost of computing quantities that depend on the commutator scales as $O(N)$ for an $N \times N$ lattice. As the highest-order coefficients of the RER also depend on the commutator (see Section 1.4.1), those also scale as $O(N)$. We can take advantage of the knowledge of the scaling by defining a per-particle RER (pp-RER). That is,

$$H_{\text{pp}}(Q_{\Delta t}|P_{\Delta t}) := \frac{1}{N} H(Q_{\Delta t}|P_{\Delta t}). \tag{1.54}$$

This way, setting a tolerance for the pp-RER will have the same meaning across different system sizes. We confirmed that $O(N)$ is the right scaling of the pp-

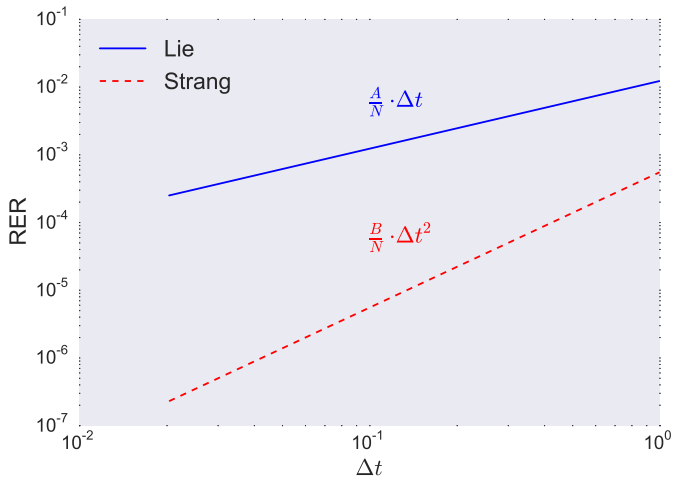


**Figure 1.3.** Logarithmic scale: Comparison between $\Delta t$ and the estimate of the pp-RER for Lie and Strang. Estimates for the constants $A, B$ come from the simulation of a $2$D Ising model on a $100 \times 100$ lattice with final time $T = 1000$. Simulation was done in parallel with SPPARKS.

RER with respect to system size via simulation, as we saw that for increasing $N$, $H_{\text{pp}}(Q_{\Delta t}|P_{\Delta t}) \simeq o(1)$.

To estimate the top-order coefficients of the pp-RER expansion, we simulated the system until convergence to the stationary distribution was established. After that, every sample simulated by SPPARKS [56] was used to calculate the estimates. Note that, in this case, we show an overestimate of $B$, so results for the Strang splitting will be even better than the ones presented in Figure 1.3. It is possible to get an estimator that converges to the exact value of $B$ by adding all of the positive terms in $L_S^3[\delta'_\sigma](\sigma)$ to the denominator of (A.5). Figure 1.3 illustrates the difference in long-time accuracy between the two splittings. Since this is a logarithmic plot, most of the difference is made by Strang having a different order than Lie.

**Remark 8** (on the efficiency of computing the highest-order coefficients of the expansion of the RER for the Lie and Strang operator splittings.)**.** In the case of a checkerboard decomposition of the lattice (see Figure 1.1), we can calculate in exactly

**Table 1.1.** Upper bounds (normalized by lattice size) on the number of lattice sites we need to evaluate the transition rates at in order to calculate the commutator for each operator splitting, assuming that a checkerboard decomposition into $m^2$ sublattices of an $N \times N$ lattice is used, as in Figure *1.1*. The commutator also encodes the cost of communication between the processes. As $N$ grows, the cost of communication is smaller, as the processes spend more time simulating on the sublattices than updating each others' boundaries.

| | Lie | Strang |
|---|---|---|
| Upper bound of the commutator cost (normalized by number of sites, $N^2$) | $2(m+1)/N$ | $6(m+1)/N$ |

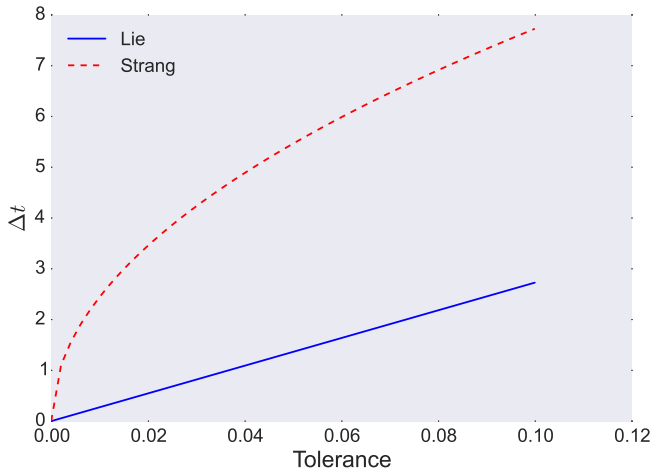


**Figure 1.4.** Comparison between tolerance and $\Delta t$. The difference in order of the pp-RER between the two splittings allows for a larger splitting time step $\Delta t$ given a fixed tolerance. This is similar to the behavior of the error in [4], although the RER allows us to make this statement for $T \gg 1$.

how many sites we need to evaluate the rates in order to calculate the commutator. However, for our purposes, upper bounds will be more appropriate. Table 1.1 offers a comparison of those bounds when we decompose an $N \times N$ lattice into $m^2$ sublattices, assuming nearest neighbor interactions. Notice that the cost is larger for Strang due to the complexity of the corresponding commutator.

On a more practical note, a user of a splitting scheme may instead like to see the flipped relationship. That is, given a fixed tolerance, what is the maximum time window during which the simulation can run asynchronously? If we interpret

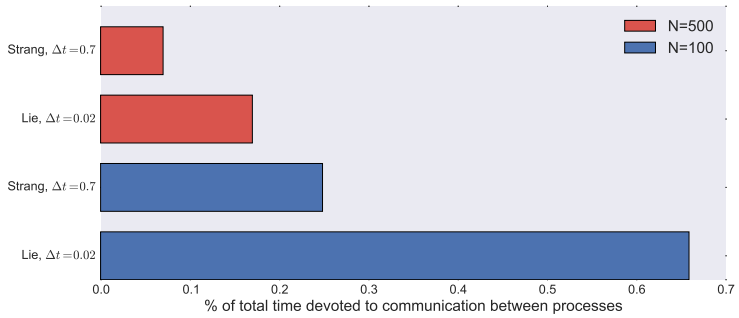


**Figure 1.5.** Percentage of time each scheme devotes to communication in a fixed time interval, $[0, T]$, for a square $N \times N$ lattice when simulating an Ising-type system, using four processes and for $T = 3000$. Note that for the $\Delta t$ considered, the pp-RER tolerance is $10^{-3}$ for both schemes. Due to the considerably smaller step size of the Lie scheme, a larger chunk of time is devoted to communication. This is more apparent in the case of a moderately small lattice, $N = 100$, where the time spent updating the other processes is over 60% of total time. Communication cost is more severe when $N$ is smaller. By Remark *8*, as $N$ grows, communication should take less of the total time, as the processes spent more time simulating than updating their boundaries.

tolerance as a fixed value of $H_{\text{pp}}(Q_{\Delta t}|P_{\Delta t})$ during the simulation, then the relationship with $\Delta t$ is the one in Figure 1.4. There we can see that if our error tolerance with respect to the pp-RER is $10^{-3}$, then any $\Delta t$ smaller than 0.7 works for the Strang splitting. To get within the same tolerance with Lie, $\Delta t$ has to be less than 0.02, a substantially smaller step-size for parallel computations. As is expected, a smaller step-size comes with larger communication cost and thus a longer computation for the same tolerance. This can be seen in Figure 1.5.

**Remark 9.** Figures 1.4 and 1.5 illustrate the very practical consequences of the theory. Interest in highly accurate splitting schemes in PL-KMC stems from a tolerance-versus-communication point of view. A user of such a scheme would like for it to be as accurate as possible; therefore the step size, $\Delta t$, should be relatively small. However, for the scheme to be efficient, $\Delta t$ should be large enough for every processor to have a substantial amount of work to do before communications are in order. A good balance can be reached in between and a scheme that is more accurate allows

for a larger $\Delta t$ while holding the same error tolerance. Given that the RER captures long-time behavior, this is an important comparison between the schemes.

### 1.5.1 The pp-RER as an efficient diagnostic quantity for parallel KMC

The discussion above about the pp-RER, (1.54), suggests the use of these estimates as efficient diagnostic quantities for comparing schemes. As discussed in the previous section, we can infer the scaling of the top-order coefficient of the RER by the properties of the commutator. Consequently, we can "normalize" the RER (as in (1.54)) by that scaling to derive a similarity measure that does not depend on system size. This is significant as it allows practitioners to compare schemes and tune parameters ($\Delta t$, domain decomposition, etc.) on a system of smaller size and thus avoid further slowing down of the target simulation, which is crucial for complicated systems. Overall, our approach can be viewed as a diagnostic tool that allows us to compare different parallelization schemes based on operator splitting.

## 1.6 Some connections with model selection and information criteria

The interacting particle system application considered in Section 1.5 allows us to look at the RER via a statistical lens. The goal is to compare two models, $Q^1_{\Delta t}, Q^2_{\Delta t}$, of the actual distribution $P_{\Delta t}$ by utilizing simulated data. From this standpoint, our methodology is nothing more than model selection. There is an abundance of literature toward tackling the comparison of different models, given a sufficiently large amount of data. A prominent example is the use of information criteria in the model selection literature, like Akaike [2] and Bayesian [3]. Those provide estimates for the information lost compared to a given data set by using one approximate model instead of another, without requiring knowledge of the true model.

The approach in this work is very similar in nature. As stated before, motivated by Theorem 7, we can express the RER in each case as

$$H(Q^i_{\Delta t}|P_{\Delta t}) = A_i \Delta t^{p_i} + o(\Delta t^{p_i}), p_i \geq 1, i \in \{1, 2\}.$$

For instance, in the case of the Lie splitting, $A_1 = A$ as defined in (A.1), $p_1 = 2$, and for Strang $A_2 = B, p_2 = 3$, as defined in (A.3). Given simulated data and for a small fixed $\Delta t$, we can estimate the coefficients $A_i$. Comparison of the schemes can now be done through

$$H(Q^1_{\Delta t}|P_{\Delta t}) - H(Q^2_{\Delta t}|P_{\Delta t}) = A_1 \Delta t^{p_1} - A_2 \Delta t^{p_2} + o\left(\Delta t^{\min(p_1,p_2)}\right). \tag{1.55}$$

The difference $A_1 \Delta t^{p_1} - A_2 \Delta t^{p_2}$ shares the properties of the information criteria previously mentioned while also introducing some new ones:

1. It is a computationally tractable quantity.
2. It compares the schemes in terms of long-time information loss (through $p_1, p_2$).
3. It takes into account communication cost of each scheme (through $A_1, A_2$ and associated commutators).

Thus, as an information criterion, RER differences like in (1.55) offer a different perspective through which to pick a splitting scheme over another. A new element in our approach, compared to the earlier vast literature on information criteria, is the use of RER instead of the standard relative entropy. Using RER allows us to compare stochastic dynamics models and in a data context, correlated time series.

## 1.7 Generalizations, connectivity, and relative entropy rate

Up to this point, we have analyzed the RER with respect to the leading order in $\Delta t$ for the case of a stochastic particle system (see Theorem 7). In this section, we

study the RER in a more general setting and illustrate that it captures important details about the system and the scheme. We will also see how the order of the RER can change depending on those details, resulting in some cases in schemes of higher accuracy. We showcase this with a simple Markov chain example in Section 1.7.1.

**Definition 10** (restriction of a generator)**.** *Let us have a set $A$ with $A \subset S \times S$ and $L$ be an infinitesimal generator of a Markov process with associated transition rates $q$. Then, the restriction $L|_A$ of $L$ is defined as*

$$L|_A[f](\sigma) = \sum_{\sigma' \in S} q_A(\sigma, \sigma') \left( f(\sigma') - f(\sigma) \right), \ \sigma \in S, \tag{1.56}$$

*where $q_A(\sigma, \sigma') = q(\sigma, \sigma') \cdot \chi_A(\sigma, \sigma')$, $\chi_A$ is the characteristic function of set $A$, and $f$ is a continuous and bounded function on the state space $S$.*

We assume that the operator $L$ is split into $L_1, L_2$ and that both are *restrictions* of $L$. Note that Definition 10 is general enough to include the splittings used in PL-KMC. For example, the generators $L_1, L_2$ in (1.18) are precisely of that form, with the groups $G_i$ playing the role of the sets "$A$."

Before we can construct an asymptotic estimate for the RER, we need to first introduce some of the tools we will use. Let $\sigma, \sigma'$ be states of a CTMC on a countable state space and let $q$ be the associated transition rates. Then, a path $\vec{z} = (z_0, \ldots, z_n)$ from $\sigma$ to $\sigma'$ is a finite sequence of distinct states $z_i$ such that $z_0 = \sigma, z_n = \sigma'$, and $\prod_{i=0}^{n} q(z_i, z_{i+1}) > 0$. The length of a path will be denoted by $|\vec{z}| = |(z_0, \ldots, z_n)| = n$ and we will use $\mathrm{Path}(\sigma \to \sigma')$ for the set of all paths from $\sigma$ to $\sigma'$. Thus, we are now able to define a distance between states by looking at the length of the shortest path that connects them.

**Definition 11** (distance between states)**.** *Let $q$ be the transition rates of a CTMP over a countable state space $S$. Then, let $\sigma, \sigma' \in S$, $\sigma \neq \sigma'$. The distance $d_q$ between the two states is defined as*

$$d_q(\sigma, \sigma') := \min \{|\vec{z}| : \vec{z} \in \mathrm{Path}(\sigma \to \sigma')\} . \tag{1.57}$$

*In the case that the two states are disconnected, i.e.,* $\mathrm{Path}(\sigma \to \sigma') = \emptyset$*, then $d(\sigma, \sigma') = +\infty$. Given those distances, one can also define the diameter of the space as*

$$\mathrm{diam}(S) = \max_{(\sigma,\sigma') \in S \times S} \{d(\sigma, \sigma')\}.$$

This notion of distance comes from graph theory and is known as the geodesic distance. When there is no ambiguity concerning the transition rates used, we will drop the $q$ from the notation, using $d$ instead of $d_q$. $d$ is not a metric in the classical sense, since it does not have to be symmetric, that is, $d(\sigma, \sigma') \neq d(\sigma', \sigma)$ in general. However, it satisfies the triangle inequality. In addition, the distances depend only on the transition rates, i.e., they are time independent. We will refer to those distances as the *connectivity* of the state space for the Markov chain with transition rates $q$. The importance of using such a distance can be seen in the following result concerning compositions of the infinitesimal generator $L$.

**Lemma 12.** *Let $L$ be an infinitesimal generator of a Markov process, with corresponding transition rates $q$, and let $\sigma'$ be some state of the process. Then,*

$$\{\sigma : L^n[\delta_{\sigma'}](\sigma) \neq 0\} \subseteq \{\sigma : d(\sigma, \sigma') \leq n\} = B_n(\sigma').$$

*Proof.* The proof is by induction. The argument can be found in the appendix. □

In other words, for a fixed state $\sigma'$, if $d(\sigma, \sigma') > n$, then $L^n[\delta'_\sigma](\sigma) = 0$. The set $B_n(\sigma')$ contains all states that are connected with $\sigma'$ with $n-2$ or less in between states. We will also use the notation $S_n(\sigma') := \{\sigma : d(\sigma, \sigma') = n\}$.

Since our primary interest is in studying approximations based on splitting our generator $L$ to $L_1, L_2$, it makes sense to have an extension of the previous result to compositions of $L_1, L_2$. The following lemma is the generalization of Lemma 12 to compositions of restrictions. We will use the notation $L^k|_A$ to denote the $k$th composition of generator $L$, where, instead of the original transition rates, we use $q_A$.

**Lemma 13.** *Let us have the state space $S$ and $S \times S = A \cup B, A \cap B = \emptyset$, along with generators $L_1 = L|_A$, $L_2 = L|_B$. We fix $\sigma' \in S$ and $k, m \in \mathbb{N}$. Then,*

$$\left\{\sigma : L_1^k \left[L_2^m[\delta_{\sigma'}]\right](\sigma) \neq 0\right\} \subseteq \{\sigma : d(\sigma, \sigma') \leq k + m\}.$$

*Proof.* The proof is an induction argument similar to that of Lemma 12; see supplementary materials in the appendix. □

Lemma 13 can be simply extended to more complicated compositions by the use of similar arguments. Thus, if every composition of $L_1, L_2$ is controlled in the sense of Lemma 13, then it is not difficult to see that the same control holds for collections of them of the same order, i.e., if we fix $\sigma' \in S$ and $k \in \mathbb{N}$,

$$\{\sigma : L_Q^k[\delta'_\sigma](\sigma)\} \subseteq \{\sigma : d(\sigma, \sigma') < k\}. \tag{1.58}$$

We can use restrictions of generators as building blocks for splitting schemes. A point often made in this work is the importance of the commutator in studying those schemes. Thus, it makes sense to have a relation between connectivity and the commutator.

**Lemma 14** (support of the commutator)**.** *Let $L$ be the generator of a Markov process and $L_1, L_2$ restrictions of that generator. Let also $\Delta t > 0$. Then, assume $Q_{\Delta t}$ is an approximation of $P_{\Delta t}$ by using a splitting scheme of order $p$ with associated commutator $C$. Then, for fixed $\sigma' \in S$,*

$$\{\sigma : C(\sigma, \sigma') \neq 0\} \subseteq \{\sigma : d(\sigma, \sigma') \leq p\}.$$

*Proof.* In Lemma 2, we defined the commutator as $C(\sigma, \sigma') = (L^p - L_Q^p)\delta_{\sigma'}(\sigma)$. From Lemma 12, we have that if $d(\sigma, \sigma') > p$, then $L^p[\delta'_\sigma](\sigma) = 0$ and from (1.58), $L_Q^p[\delta'_\sigma](\sigma) = 0$. This gives the result. □

When the state space is finite, as in the case of stochastic particle systems on finite lattices, then the commutator $C$ is a matrix indexed by the different states. An implication of Lemma 14 is that there is a reordering of the rows/columns that turns $C$ into a banded matrix. Regardless, we can now prove a general result for the asymptotics of the RER.

**Theorem 15.** *Consider $\Delta t \in (0, 1)$ and let $P_{\Delta t}(\sigma, \sigma') = e^{L\Delta t}\delta_{\sigma'}(\sigma)$, $Q_{\Delta t}(\sigma, \sigma')$ be an approximation of $P_{\Delta t}$ based on a splitting scheme with $L_1, L_2$ restrictions of the generator $L$ and $\mu_{\Delta t}$ the stationary measure corresponding to $Q_{\Delta t}$. Then, if the splitting scheme is of order $p$, we define the bounded diameter of the state space as $\hat{k}$,*

$$\hat{k} = \min\{\text{diam}(S), p\} = \min\{\max_{\sigma,\sigma'}\{d(\sigma, \sigma')\}, p\}.$$

*Then, if $C(\sigma, \sigma') \neq 0$ for at least one pair $\sigma, \sigma' \in S$ such that $d(\sigma, \sigma') = \hat{k}$, we have that*

$$H(Q_{\Delta t}|P_{\Delta t}) = O(\Delta t^{2p-(\hat{k}+1)}).$$

*Proof.* The proof of this theorem is the generalization of the argument given for Theorem 7. Picking up from formula (1.45),

$$J(\Delta t;\sigma,\sigma') = \frac{(C(\sigma,\sigma'))^2}{2Q_{\Delta t}(\sigma,\sigma') + \Delta t^p C(\sigma,\sigma') + o(\Delta t^p)}\Delta t^{2p} + o(\Delta t^{2p-\hat{k}}). \tag{1.59}$$

Our goal is to show that $J(\Delta t;\sigma,\sigma') = O(\Delta t^{2p-\hat{k}})$ for some $(\sigma,\sigma')$ and that this is the highest order attainable. Next, let us have $(\sigma,\sigma') \in S \times S$ such that $d(\sigma,\sigma') = \hat{k}$. Then, from (1.11) and (1.58), we have that

$$Q_{\Delta t}(\sigma,\sigma') = \sum_{k=\hat{k}}^{\infty} \frac{L_Q^k[\delta_{\sigma'}](\sigma)}{k!}\Delta t^k = O(\Delta t^{\hat{k}}),\ \Delta t \in (0,1]. \tag{1.60}$$

Thus from (1.59) and (1.60), we can expose the first term of the asymptotic expansion of $F$ as

$$J(\Delta t;\sigma,\sigma') = \begin{cases} \dfrac{(C(\sigma,\sigma'))^2}{2L_Q^{\hat{k}}[\delta_{\sigma'}](\sigma)/k!}\Delta t^{2p-\hat{k}} + o(\Delta t^{2p-\hat{k}}), & \hat{k} < p, \\ \dfrac{(C(\sigma,\sigma'))^2}{2L_Q^{\hat{k}}[\delta_{\sigma'}](\sigma)/k! + C(\sigma,\sigma')}\Delta t^p + o(\Delta t^p), & \hat{k} = p. \end{cases} \tag{1.61}$$

Next, we need to address the contribution of the rest of the expansion used (see the proof of Theorem 7), that is,

$$G(\Delta t;\sigma,\sigma') = Q_{\Delta t}(\sigma,\sigma')\sum_{k=1}^{\infty}\frac{1}{2k+1}\left(\frac{Q_{\Delta t}(\sigma,\sigma') - P_{\Delta t}(\sigma,\sigma')}{Q_{\Delta t}(\sigma,\sigma') + P_{\Delta t}(\sigma,\sigma')}\right)^{2k+1}.$$

If $\hat{k} < p$, then $G(\Delta t;\sigma,\sigma') = O(\Delta t^{3p-2\hat{k}})$, which are lower-order terms given that $\Delta t \leq 1$. However, if $\hat{k} = p$, $G(\Delta t;\sigma,\sigma') = O(\Delta t^p)$ and in fact every term of the series in $G$ is of that order.

Finally, $H(Q_{\Delta t}|P_{\Delta t})$ can never have higher order than $p-1$, as that would require $(\sigma,\sigma')$ such that $d(\sigma,\sigma') > p+1$ and then $C(\sigma,\sigma') = 0$ (from Lemma 14). □

The assumption on the commutator in Theorem 15 is simple to check for parallel KMC, as we can write down the commutator $C(\sigma, \sigma')$ explicitly. For example, for Lie, $C(\sigma, \sigma')$ is given by (1.34), so checking the assumption is just a matter of calculation. Additionally, to find the bounded diameter $\hat{k} = \min\{\text{diam}(S), p\}$, it is sufficient to have lower bounds for the diameter, $\text{diam}(S)$, as the order of the local error of the scheme, $p$, will typically be much smaller. Example 1.7.1 shows a case where $p$ is close to $\text{diam}(S)$ and the implications this has for the RER.

### 1.7.1 Markov chain example

In order to illustrate the connectivity-RER relation, we are studying a simple example where we can compute the RER and all related quantities explicitly, either by hand or by any symbolic algebra system. All calculations of the RER in this example are not from sampling but by using definition (1.22).

We study the case of a Markov process with transition rate matrix, $Q$ and $\text{diam}(S) = 2$. We consider a positive $\Delta t$, $\Delta t < 1$, and

$$Q = \begin{pmatrix} -3 & 1 & 2 \\ 3 & -4 & 1 \\ 1 & 0 & -1 \end{pmatrix}.$$

Given this, we can calculate the transition probability matrix of the Markov chain as the matrix exponential of $Q$, $P_{\Delta t}(\sigma, \sigma') = \exp(\Delta t Q)\delta_{\sigma'}(\sigma)$. Our system has diameter equal to two since $Q_{3,2} = 0$ but $Q_{3,1} \cdot Q_{1,2} \neq 0$. We can construct approximations of $P_{\Delta t}$ by splitting $Q$ into components $A, B$ with $Q = A + B$, similarly to how we expressed the generator $L$ as $L_1 + L_2$. One way to do this is

$$A = \begin{pmatrix} -3 & 1 & 2 \\ 3 & -4 & 1 \\ 0 & 0 & 0 \end{pmatrix}, B = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & 0 & -1 \end{pmatrix}.$$

Thus, one approximation of $\exp(Q\Delta t)$ could be $\exp(A\Delta t)\exp(B\Delta t)$, which corresponds to the Lie splitting. From Theorem 15, since $\mathrm{diam}(S) = p = 2$, we expect $H(Q^{\mathrm{Lie}}_{\Delta t}|P_{\Delta t}) = O(\Delta t^1)$. This is indeed the case, as

$$H(Q^{\mathrm{Lie}}_{\Delta t}|P_{\Delta t}) \simeq 0.124\Delta t - 0.0566\Delta t^2 + O\left(\Delta t^3\right).$$

The use of $\simeq$ comes from a truncation of the coefficients to three significant digits. We can work similarly with the Strang splitting, now using $\exp(A\Delta t/2)\exp(B\Delta t)\cdot \exp(A\Delta t/2)$ as the approximation to $P_{\Delta t}$. The local order of the Strang splitting is $p = 3$, so we expect that $H(Q^{\mathrm{Strang}}_{\Delta t}|P_{\Delta t}) = O(\Delta t^{2\cdot 3-3}) = O(\Delta t^3)$ (see Theorem 15). This can be readily demonstrated by a calculation of the RER, followed by the derivation of its asymptotic expansion:

$$H(Q^{\mathrm{Strang}}_{\Delta t}|P_{\Delta t}) \simeq 0.0279\Delta t^3 + 0.000672\Delta t^4 + O\left(\Delta t^5\right).$$

## 1.8 Quantifying information loss in transient regimes

In this last section, we consider the case where we wish to study the performance of the operator splitting scheme in a transient regime, before convergence to the stationary distribution takes place. Note that in the proofs of Theorems 7 and 15, we derived the asymptotic expressions of the various quantities without referring to the stationary measure $\mu_{\Delta t}$. Therefore those results do not depend on the choice of the sampling measure. That is, with the assumptions of Theorem 15 and $\nu$ a probability distribution on the state space $S^M$ such that $\nu(\sigma) > 0$ for all states $\sigma$, then

$$H_\nu(Q_{\Delta t}|P_{\Delta t}) = \sum_{\sigma\in S^M} \nu(\sigma) Q_{\Delta t}(\sigma_0,\sigma_1)\frac{Q_{\Delta t}(\sigma_0,\sigma_1)}{P_{\Delta t}(\sigma_0,\sigma_1)} = O(\Delta t^{2p-\hat{k}}). \tag{1.62}$$

Therefore, the order of the RER is independent of the sampling measure. As a result, we gain Theorem 16, an extension of Theorem 15 to transient time regimes.

**Theorem 16.** *With the assumptions of Theorem* 15 *for the RER, we have that for any* $T > 0$

$$\frac{R(Q_{0:T}|P_{0:T})}{T} = \frac{R(\mu_0|\nu_0)}{T} + O\left(\Delta t^{2p-\hat{k}}\right). \tag{1.63}$$

Theorem 16 is implied by the decomposition of the relative entropy in terms of rates that depend on $\nu_i$ (first discussed in Section 1.2). If $M$ is a positive integer, $\Delta t$ is the scheme's time step, and $T = M\Delta t$, then

$$R(Q_{0:T}|P_{0:T}) = R(\mu_0|\nu_0) + \sum_{i=1}^{M} H_{\nu_i}(Q_{\Delta t}|P_{\Delta t}). \tag{1.64}$$

*Proof of Theorem* 16. From (1.62) we have that the order of the RER does not depend on the sampling measure $\nu$, as long as $\nu(\sigma) > 0$ for all $\sigma$. Therefore, $H_{\nu_i}(Q_{\Delta t}|P_{\Delta t}) = O(\Delta t^{2p-\hat{k}})$ for $i = 1, \ldots, M$. This, combined with (1.64), implies the result. □

Therefore, our results about the RER are applicable for parallel KMC even for practitioners that are interested in simulating the dynamics in the transient regime.

**Remark 17** (RER versus pathwise relative entropy)**.** In Section 1.2, we saw that, in the stationary regime, we can relate the pathwise relative entropy with the RER via

$$R(Q_{0:T}|P_{0:T}) = TH(Q_{\Delta t}|P_{\Delta t}) + R(Q_{\Delta t}|P_{\Delta t}).$$

In this section, we connected the RER with the relative entropy for transient regimes by using relation (1.64). Ultimately, those relations motivate the use of the RER as an information criterion in place of the pathwise relative entropy, but there are other advantages too:

1. The RER does not depend on the length of the simulated path. Additionally, it can be estimated from a single path, while the pathwise relative entropy requires several.

2. For large $T$, the relative entropy and RER encapsulate the same amount of information about the similarity of $Q_{\Delta t}$ and $P_{\Delta t}$.

## 1.9 Conclusions

We introduced the RER, i.e., path-space relative entropy per unit time, as a means to quantify the long-time accuracy of splitting schemes for stochastic dynamics and in particular parallel KMC algorithms. We demonstrated, using a posterirori error expansions, the dependence of RER on the following elements: the local error analysis of the splitting schemes captured by the operator commutators; the local error order $p$ and the splitting time step $\Delta t$, which in the case of Parallel KMC controls the asynchrony between processors; and the diameter of the graph associated with the approximated Markov jump process.

Based on this analysis, we showed that RER defines a computable path-space information criterion that allows us to compare, select (and possibly design) different splitting schemes, taking into account both error tolerance (e.g., accuracy of the scheme) and practical concerns such as asynchrony and processor communication cost. It is also appropriate to think of the RER as a diagnostic quantity that can be estimated on systems of smaller size and consequently be used to compare schemes and tune parameters without slowing down the target simulation.

Finally we note that numerical analysis of stochastic systems is typically concerned with controlling the weak error for observable functions $\phi$,

$$\sup_{0\leq n\leq N} \left|\mathbb{E}_{P_{0:T}}[\phi(X(n\Delta t))] - \mathbb{E}_{Q_{0:T}}[\phi(X_n)]\right|, \tag{1.65}$$

where $X_n$ represents the approximate chain and $X(n\Delta t)$ the $\Delta t$-skeleton chain of the exact process, $T = M \cdot \Delta t$. However, our results measure the information loss on path space between the approximate chain and the $\Delta t$-skeleton chain of the exact process, using RER. Controlling RER also implies upper bounds for observables at long times, using uncertainty quantification information inequalities developed in [22, 36]. We also showed how those results can be extended to finite-time regimes.

# CHAPTER 2

# INFORMATION METRICS FOR QUANTIFYING LOSS OF REVERSIBILITY IN PARALLELIZED KMC

In this chapter[1], we study operator splitting schemes for PL-KMC from the point of view of time-reversibility. Although the original Markov Process can satisfy the detailed balance condition, the approximating process simulated by PL-KMC will not. This is due to the time discretization as well as the domain decomposition and asynchronous computation required for efficient simulation. We propose the entropy production rate (EPR) as a tool that: 1. captures the loss of time-reversibility for an operator splitting scheme, 2. is an *a posteriori* quantity that can be estimated during the parallel simulation and 3. can be used to discriminate between the performance of a variety of schemes. We discuss the estimation of the EPR on an adsorption-desorption example simulated with Sandia Labs' SPPARKS code. For this example, we also compare the loss of reversibility if we change the domain decomposition from blocks to stripes. We notice that stripes tend to behave better but also cost more in terms of computer memory.

## 2.1 Background on Parallel Lattice KMC

Parallel Lattice KMC is an approximation to the exact, but serial, simulation algorithm. In implementations, it works by taking advantage of the spatial dependencies between the different events. For example, in a model with finite range interactions, the spins on two lattice sites can change with no error to the dynamics as long as the

[1]The contents of this chapter are published in the Journal of Computational Physics [28] and appear here with permission.

two are sufficiently far apart. Therefore, by decomposing the lattice into sub-lattices, we gain an efficient alternative to serial KMC analogous to domain decomposition methods in parallel algorithms for partial differential equations.

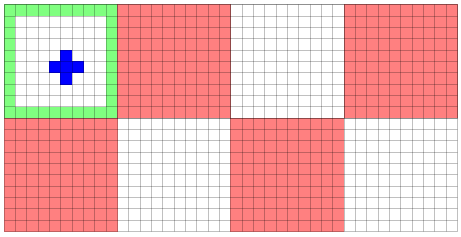

**Figure 2.1.** Checkerboard decomposition of a rectangular lattice into sub-lattices. Because each site's transition depends on the information from the nearest neighbors, transitions in sub-lattices of the same color are independent. White sub-lattices can be simulated asynchronously in time, while keeping the states in the red ones **frozen**. When the stochastic time reaches $\Delta t$, information is shared with the red sub-lattices about the state of the boundary regions (here only shown for the first sub-lattice).

A new insight provided in [5] was that parallel algorithms, such as the one described in Figure 2.1, can be formulated as operator splitting schemes. This connection allows for the design, error quantification, and performance analysis of such algorithms [4]. Specifically, this approach allows for an observable-focused error analysis, through which a practitioner can pick both the scheme class and specific parameters that fit the computational needs. Additionally, it formalizes the dependence of the error on the decomposition of the lattice and on the splitting time step, $\Delta t$, for bounded time intervals. Finally, it also allows to study the long-time behavior of the schemes and provides long-time error control in the recent work [26].

To begin, we pick a positive *operator splitting time step* $\Delta t$. If we were to simulate a Continuous Time Markov Chain (CTMC) via the serial KMC algorithm, then the corresponding transition probability of the process jumping from a state $\sigma$ to a state $\sigma'$, $\sigma, \sigma' \in S$, in time $t$ would be

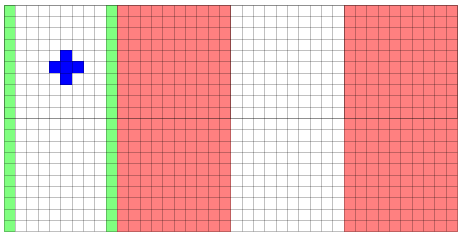

**Figure 2.2.** Stripe decomposition of a rectangular lattice into sub-lattices. Compared to Figure 2.1, now each processor needs to store more information before the runs can take place. However, if we fix the width of the blocks, then the boundary regions (here only shown for the first sub-lattice) will shrink, can lead to less error per time step [4, 26]. Considering a block decomposition with smaller block width is possible, but there are limits to how small the width can be while still preserving the efficiency of the parallel algorithm [4]. This can also be seen in Figure 2.3.

$$P_t(\sigma, \sigma') = P(\sigma_t = \sigma' | \sigma_0 = \sigma) = e^{tL}\delta_{\sigma'}(\sigma). \tag{2.1}$$

In (2.1), $\delta_{\sigma'}$ is a Dirac probability measure, centered at state $\sigma'$, and $L$ is the generator of the process which, for bounded and continuous functions $f$, is defined as

$$L[f](\sigma) := \sum_{\sigma' \in S} q(\sigma, \sigma')(f(\sigma') - f(\sigma)). \tag{2.2}$$

The transition rates of the CTMC will be denoted by $q(\cdot, \cdot)$. In general, they are tied to the system being modelled and are assumed to be known, see Appendix B.4 and B.5.

Since the approximate process will be a discretization with $\Delta t$ step size, we will be comparing it against the $\Delta t$-skeleton of the exact Continuous Time Markov Chain, with transition probability $P_{\Delta t}(\sigma, \sigma') = e^{\Delta t L}\delta_{\sigma'}(\sigma)$. This is only done to simplify the comparison and corresponds to sub-sampling the exact KMC, keeping only the states every $\Delta t$ apart. Now, inspired by the Trotter product formula [60], i.e.,

$$e^{\Delta t L} = \lim_{n\to\infty} \left( e^{\Delta t/nL_1} e^{\Delta t/nL_2} \right),$$

we can write approximations to $e^{\Delta t L}$ by splitting the operator $L$ into $L_1 + L_2$ (with associated rates $q_1, q_2$). For example, two popular approximations are:

$$e^{\Delta t L} \simeq e^{\Delta t L_1} e^{\Delta t L_2}, \qquad \text{(Lie)} \tag{2.3}$$

$$e^{\Delta t L} \simeq e^{\Delta t/2L_1} e^{\Delta t L_2} e^{\Delta t/2L_1}. \qquad \text{(Strang)} \tag{2.4}$$

Throughout this work, we shall be using $Q_{\Delta t}$ to denote the transition probability arising from approximations to $e^{\Delta t L}$. We will also use $\mu_{\Delta t}$ to denote the corresponding stationary measure.

Although we consider a splitting into two operators, $L_1, L_2$, this is for the convenience of the reader. Occasionally, it is beneficial to split the generator $L$ into more than two parts, as is done in Sandia Labb's SPPARKS code [56], where a $2D$ simulation decomposes the lattice into four pieces instead of two. However, the error analysis extends naturally to this case.

### 2.1.1 Local Error Analysis

Operator splitting approximations are equivalent to specific computational schedules for Parallel Lattice KMC schemes [5]. For example, if we alternate between the red and white groups in Figure 2.1, allowing each group to run only for $\Delta t$, then that is equivalent to using the Lie splitting (Equation (2.3)) to approximate $e^{\Delta t L}$. If $L$ is a bounded operator, then we can write the semigroup as a series expansion,

$$e^{\Delta t L} = \sum_{k=0}^{\infty} \frac{\Delta t^k}{k!} L^k, \tag{2.5}$$

where $L^k$ stands for the resulting operator after $k$ compositions of $L$. We can also write a representation for the various operator splitting schemes. For example, for the case of the Lie splitting in (2.3) and by using the expansion in (2.5),

$$\begin{aligned} e^{\Delta t L_1} e^{\Delta t L_2} &= \left(I + \Delta t L_1 + O(\Delta t^3)\right) \cdot \left(I + \Delta t L_2 + O(\Delta t^3)\right) \\ &= I + \Delta t L + \frac{\Delta t^2}{2}\left(L_1^2 + L_2^2 + 2L_1 L_2\right) + O(\Delta t^3) \end{aligned} \tag{2.6}$$

Then, the representations in (2.5) and (2.6) allow us to study the local error between $P_{\Delta t}$ and $Q_{\Delta t}$:

$$P_{\Delta t}(\sigma, \sigma') - Q_{\Delta t}(\sigma, \sigma) = \frac{\Delta t^2}{2}[L_1, L_2]\delta_{\sigma'}(\sigma) + O(\Delta t^3), \tag{2.7}$$

where $[L_1, L_2]$ is the *Lie bracket* of $L_1$, $L_2$, and is equal to $L_1 L_2 - L_2 L_1$. Similarly, the order of the local error $p$ is equal to 2. Note that $L_1, L_2$ can be expressed in terms of the transition rates, which implies that $[L_1, L_2]$ is computable for any pair of states $(\sigma, \sigma')$. A generalization of this idea is in Lemma 18.

**Lemma 18** (Commutator and Order of Local Error)**.** *Let $\sigma, \sigma'$ be states, $P_{\Delta t}$ as in Equation* (2.1) *and $Q_{\Delta t}$: approximation of $P_{\Delta t}$ via a splitting scheme. Then, there is a function $C : S \times S \to \mathbb{R}$ and an integer $p$, $p > 1$, such that*

$$P_{\Delta t}(\sigma, \sigma') - Q_{\Delta t}(\sigma, \sigma') = C(\sigma, \sigma')\Delta t^p + O(\Delta t^{p+1}). \tag{2.8}$$

*$C$ will be called the* ***commutator*** *and $p$ is the* ***order of the local error****.*

Equation (2.8) can be derived from the power series representations of $P_{\Delta t}, Q_{\Delta t}$, as $L$ is a bounded operator.

In the context of Parallel KMC, the commutator term $C = C(\sigma, \sigma')$ captures the error due to mismatches on the boundary regions between the different sub-lattices [4].

The operating assumption in this work is that all operators are bounded. This allows us to represent the transition probabilities with power series and, subsequently, to calculate the form of the commutators and of other quantities of interest (see discussion in B.3). However, the present work could also be extended to the case of unbounded operators [29, 33], where alternative representations for the semigroups could be used for the error analysis. We are not handling such cases here, as the Markov generators of stochastic particle systems are bounded operators [4].

## 2.2 Entropy Production Rate: an information criterion for reversibility

Let us consider a discrete stochastic process $X_n$, $n \in \mathbb{N}$. Then, $X_n$ is time-reversible if, for any $m \in \mathbb{N}$

$$p(\sigma_0, \ldots, \sigma_m) = p(\sigma_m, \ldots, \sigma_0), \tag{2.9}$$

where $p(\sigma_0, \ldots, \sigma_m) = p(X_0 = \sigma_0, \ldots, X_m = \sigma_0)$, $\sigma_i$ being states of the process. For stationary Markov processes, the detailed balance condition (DB) is equivalent to time-reversibility [38, Theorem 1.2]. If $X_n$ has transition probability $P$ and stationary distribution $\mu$, then DB requires that for all states $\sigma, \sigma' \in S$,

$$\mu(\sigma)P(\sigma, \sigma') = \mu(\sigma')P(\sigma', \sigma). \tag{2.10}$$

Although the DB condition (2.10) is a useful analytical tool for the construction of Markov Chains with a specific stationary distribution, we cannot apply it to quantify the loss of reversibility for the systems we are interested in. In our context,

$P$ corresponds to the transition probability, $Q_{\Delta t}$, of the scheme, which we do not know explicitly, and $\mu = \mu_{\Delta t}$ would be the stationary distribution associated with the scheme, which we can only access through sampling. In addition, due to the time-discretization, domain decomposition, and asynchronous simulation associated with the operator splitting scheme, we do not expect it to exactly satisfy condition (2.10). Consider for example the case numerical schemes for SDEs [35], where the approximation can completely break down reversibility. In view of this, we wish to quantify the loss of reversibility and connect it to the parameters of the scheme (lattice decomposition, computation schedule, time step $\Delta t$, etc.). Therefore, we need to look for alternative ways to assess the loss of reversibility of the scheme.

Returning to the definition of time-reversibility in (2.9) with respect to paths, we introduce an object from information theory, the entropy production (EP) associated with $P$:

$$\mathrm{EP}(P) = \sum_{\sigma_0,\ldots,\sigma_m} p(\sigma_0,\ldots,\sigma_m) \log\left(\frac{p(\sigma_0,\ldots,\sigma_m)}{p(\sigma_m,\ldots,\sigma_0)}\right), \tag{2.11}$$

with the sum in Equation (2.11) being over $S^m$, $S$ is the state space.

The EP is an example of a more general measure of similarity between distributions known as the relative entropy (RE), or Kullback-Leibler divergence [18]. Given two probability distributions, $p_1$, $p_2$, where $p_1$ is absolutely continuous with respect to $p_2$, then the RE of $p_1$ with respect to $p_2$ is defined as

$$R(p_1\|p_2) := \int \log \frac{dp_1}{dp_2} dp_1. \tag{2.12}$$

The definition in (2.12) enjoys the properties of a divergence: 1. $R(p_1\|p_2) \geq 0$ (Gibbs' inequality), 2. $R(p_1\|p_2) = 0$ if and only if $p_1 = p_2,\ p_1 -$ a.e. However, RE is not a metric in the strict sense, as it does not satisfy the triangle inequality and is not symmetric in its arguments.

From the second property of a divergence and (2.11), we can readily see that

$$\mathrm{EP}(P) = 0 \Leftrightarrow p(\sigma_0, \ldots, \sigma_m) = p(\sigma_m, \ldots, \sigma_0). \tag{2.13}$$

Therefore, if Equation (2.13) holds for all $m$, then that implies time-reversibility. It is because of this property of the EP that we will use it as a means to assess and quantify how much a scheme $Q_{\Delta t}$ destroys reversibility. This idea was originally motivated by tools in non-equilibrium statistical mechanics to understand long-time dynamics and fluctuations in associated non-equilibrium steady states [49, 50, 42, 41, 25].

Calculating the EP, even for moderate $m$, can be computationally intensive. From the definition in (2.11) we can derive an entropy rate that is independent of the path length when the initial sampling distribution is the stationary. By the Markov property, we can write the forward and backward path distributions as

$$\begin{aligned} p(\sigma_0, \ldots, \sigma_m) = &\quad \mu(\sigma_0)P(\sigma_0, \sigma_1)\cdots P(\sigma_{m-1}, \sigma_m), \\ p(\sigma_m, \ldots, \sigma_0) = &\quad \mu(\sigma_m)P(\sigma_m, \sigma_{m-1})\cdots P(\sigma_1, \sigma_0), \end{aligned} \tag{2.14}$$

where $\mu$ is the corresponding stationary distribution. Then, using (2.14) in Equation (2.11) and carrying out the calculations leads to

$$\mathrm{EP}(P) = m \cdot \sum_{\sigma_0, \sigma_1} \mu(\sigma_0)P(\sigma_0, \sigma_1) \log\left(\frac{P(\sigma_0, \sigma_1)}{P(\sigma_1, \sigma_0)}\right) = m \cdot \mathrm{EPR}(P). \tag{2.15}$$

A formal statement and proof of (2.15) can be found in Lemma 37, B.1. The entropy production rate (EPR) is defined for discrete time Markov processes as

$$\mathrm{EPR}(P) := \sum_{\sigma, \sigma'} \mu(\sigma)P(\sigma, \sigma') \log\left(\frac{P(\sigma, \sigma')}{P(\sigma', \sigma)}\right). \tag{2.16}$$

A more general definition, applicable to continuous-time Markov processes, can also be given, see [35] for an application in quantifying the loss of reversibility for numerical schemes for SDEs.

We will use the EPR to quantify the loss of reversibility of the schemes studied. Given $P$, we can estimate the EPR in (2.16) by the Gallavotti-Cohen functional (as done in [35]):

$$\mathrm{EPR}(P) = \lim_{N\to\infty} \frac{1}{N} \sum_{i=0}^{N} \log\left(\frac{P(\sigma_i, \sigma_{i+1})}{P(\sigma_{i+1}, \sigma_i)}\right), \tag{2.17}$$

where $(\sigma_i, \sigma_{i+1})$ are sampled according to $\mu(\sigma)P(\sigma, \sigma')$. The quantity on the right hand side of Equation (2.17) is thus, under suitable ergodic assumptions, an unbiased statistical estimator of the EPR, following the law of large numbers for Markov chains. For a given scheme $Q_{\Delta t}$ with stationary distribution $\mu_{\Delta t}$, Equation (2.16) thus becomes:

$$\mathrm{EPR}(Q_{\Delta t}) := \frac{1}{\Delta t} \sum_{\sigma,\sigma'} \mu_{\Delta t}(\sigma) Q_{\Delta t}(\sigma, \sigma') \log\left(\frac{Q_{\Delta t}(\sigma, \sigma')}{Q_{\Delta t}(\sigma', \sigma)}\right). \tag{2.18}$$

**Remark 19.** *In Equation* (2.18)*, we normalize with the time step* $\Delta t$ *since the EPR is a quantity defined as "per unit time" (see also Equation* (2.15)*). This normalization is also practically important, as we wish to consider comparisons of EPRs for different time-steps* $\Delta t$*. Finally, the same normalization was considered for the RER in previous work [26, Remark 4.2] and Equation* (2.22)*.*

The EP can also be seen as an information criterion for operator splitting schemes. Consider two schemes, $Q^1_{\Delta t}, Q^2_{\Delta t}$ that approximate the same exact $P_{\Delta t}$. Then we can use EP to quantify which of the two retains more reversibility per time step. That is, we are also interested in making statements of the form

$$\mathrm{EP}(Q^1_{\Delta t}) \leq \mathrm{EP}(Q^2_{\Delta t}). \tag{2.19}$$

We can then consider $EP(Q^1_{\Delta t}) - EP(Q^2_{\Delta t})$ as an information criterion that takes into account loss of reversibility, similarly to how AIC and BIC are used to assess the

quality of models in statistics [2, 3]. As the EP is a difficult quantity to compute, we can employ the EPR and Equation (2.15), and thus have another way to distinguish possible schemes based on their performance in controlling the loss of reversibility. In analogy with Inequality (2.19), we are interested in the difference

$$\mathrm{EPR}(Q^1_{\Delta t}) - \mathrm{EPR}(Q^2_{\Delta t}). \tag{2.20}$$

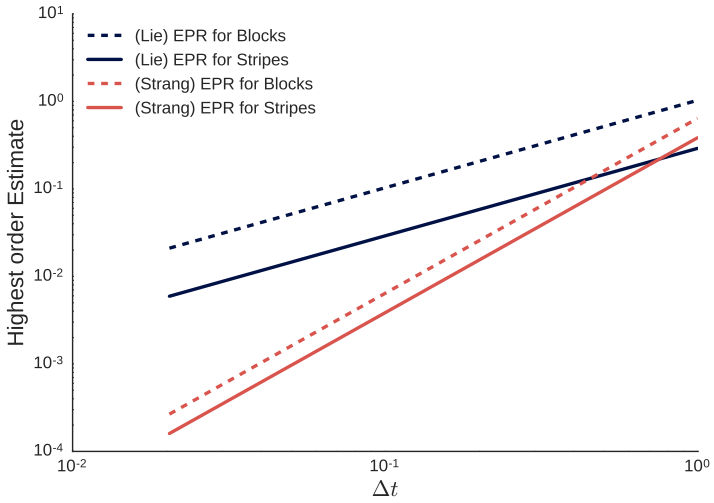


**Figure 2.3.** Approximations to the EPR of the form $(A + D) \cdot \Delta t^{p-1}$ for small $\Delta t$. The Strang scheme retains more reversibility per time step and is more "stable" (with respect to the entropy production rate) under changes in the decomposition. Also, note that the estimate is normalized by $\Delta t$ as per Remark 19. The example is an adsorption/desorption system, see B.4 for details on the system and B.3 for the estimator formulas.

Even though we have an abstract representation of $Q_{\Delta t}$ (see Equations (2.3) and (2.4)), we cannot calculate $Q_{\Delta t}$ directly. What we do know explicitly are the transition rates of the process. We can leverage this information to construct a series expansion of $Q_{\Delta t}$ around $\Delta t$ where each term depends on the transition rates. Through this, we can build statistical estimators of the highest order terms in an expansion of the EPR. Details about the coefficients and their statistical estimation

are in Sections 2.3, 2.4. In Figure 2.3 we demonstrate a comparison of two different parallel KMC schemes, based on these computable *a posteriori* expansions of EPR.

## 2.3 Loss of reversibility in Parallel KMC

In this section, we will demonstrate how to use the EPR to quantify and control the loss of reversibility for parallel Kinetic Monte Carlo (P-KMC). We will also mention details about the implementation of the various observables that are needed in order to estimate EPR.

As mentioned before, for stochastic particle dynamics, we cannot directly apply the definition in Equation (2.18), as we do not have the transition probabilities $Q_{\Delta t}$ explicitly. Instead, we will use asymptotic results to approximate the EPR for a small splitting time step, $\Delta t$ (see Section 2.4 for derivations). We first write the EPR as per Theorem 22, Section 2.4, but taking also into consideration Remark 19 for the required $\Delta t$ normalization. That is,

$$\mathrm{EPR}(Q_{\Delta t}) = H(Q_{\Delta t}|P_{\Delta t}) + I(Q_{\Delta t}|P_{\Delta t}), \tag{2.21}$$

where $H$ represents the relative entropy rate (RER)

$$H(Q_{\Delta t}|P_{\Delta t}) := \frac{1}{\Delta t}\sum_{\sigma,\sigma'} \mu_{\Delta t}(\sigma) Q_{\Delta t}(\sigma,\sigma') \log\left(\frac{Q_{\Delta t}(\sigma,\sigma')}{P_{\Delta t}(\sigma,\sigma')}\right) \tag{2.22}$$

and $I$ is a "discrepancy" term (see Section 2.4) defined as

$$I(Q_{\Delta t}|P_{\Delta t}) := \frac{1}{\Delta t}\sum_{\sigma,\sigma'} \mu_{\Delta t}(\sigma) Q_{\Delta t}(\sigma,\sigma') \log\left(\frac{P_{\Delta t}(\sigma',\sigma)}{Q_{\Delta t}(\sigma',\sigma)}\right). \tag{2.23}$$

Before we move on to how results on the RER and $I$ combine to give an asymptotic picture of the EPR, we shall first discuss what each of those captures. The RER, or

relative entropy per unit time, has been used in previous work [26] as a means to quantify the long-time error of operator splitting schemes in the context of parallel KMC. Because of this, the RER can be used as an information criterion to compare such schemes, as it takes into account details of the scheme such as the splitting time step, the domain decomposition of the lattice, and the computational schedule used. The RER has the properties of a divergence, i.e. non-negativity for any $Q_{\Delta t}, P_{\Delta t}$, and equality with zero if and only if $Q_{\Delta t} = P_{\Delta t}$. The discrepancy term in Equation (2.23) is what enforces the property of the EPR to be zero when $Q_{\Delta t}$ is time-reversible. As we shall see in Section 2.4, $I$ is not a divergence.

Now, by the individual results for the asymptotic behavior of RER (see proof of Theorem 8.6 in [26]) and $I$ (see Equation (2.41)) for small $\Delta t$, we have

$$H(Q_{\Delta t}|P_{\Delta t}) = A \cdot \Delta t^{p-1} + O(\Delta t^p), \tag{2.24}$$

$$I(Q_{\Delta t}|P_{\Delta t}) = D \cdot \Delta t^{p-1} + O(\Delta t^p). \tag{2.25}$$

Therefore, from Equations (2.21), (2.24), and (2.25), we get

$$\mathrm{EPR}(Q_{\Delta t}) = (A + D)\Delta t^{p-1} + O(\Delta t^p). \tag{2.26}$$

We remind here that $p$ stands for the order of the local error (see Lemma 18).

Coefficients $A$ and $D$ are expected values of specific observables with respect to $\mu_{\Delta t}$ (see B.4 for the explicit formulas in the case of an adsorption/desorption process and B.5 for the case of a diffusion process). Therefore, under some ergodicity assumptions, they can be estimated via simulation of the system by using the parallel algorithm. In Figures 2.3 and 2.4, we estimate the EPR by an estimation of the constants $A, D$ for small timestep $\Delta t$.

In previous work [26], we expressed $A$ explicitly in terms of the commutator $C$ and the transition rates of the original process. For example, given a lattice $\Lambda$, for

the Lie splitting and an adsorption/desorption example (see B.4), the highest order coefficient for the RER is:

$$A = A_{\text{Lie}} = \mathbb{E}_{\mu_{\text{Lie}}}\left[\sum_{x,y\in\Lambda} C_{\text{Lie}}(\sigma,\sigma^{x,y})F_{\text{Lie}}(\sigma,\sigma^{x,y})\right] \tag{2.27}$$

$$= \sum_{\sigma}\mu_{\text{Lie}}(\sigma)\sum_{x,y\in\Lambda} C_{\text{Lie}}(\sigma,\sigma^{x,y})F_{\text{Lie}}(\sigma,\sigma^{x,y}), \tag{2.28}$$

where $\mu_{\text{Lie}}$ is the corresponding stationary distribution of the Lie scheme, $C_{\text{Lie}} = [L_1, L_2]$ and $F_{\text{Lie}}$ depends only on the transition rates. If we consider a state $\sigma$ and a lattice site $x$, $\sigma^x$ corresponds to the resulting state after a spin-flip at that lattice site and $\sigma^{x,y}$ denotes successive spin-flips at $x$ and $y$. Note that $A_{\text{Lie}}$ in Equation (2.27) seemingly depends on all lattice positions $x, y$. This is also the case for $D_{\text{Lie}}$ and the corresponding coefficients for the Strang splitting (see B.3). However, an important property of the commutator in Lemma 18 can be used to simplify the situation and is further explained in Remark 20.

**Remark 20.** *A key result in [4] was that the commutator is non-zero only for lattice sites on the boundary regions (see Figure 2.1). This has two major implications:*

1. *The sums over the lattice $\Lambda$ in the highest order coefficients, A and D (see Equation (2.24) and (2.25)), are really sums over the boundary regions, as the commutators for Lie and Strang are non-zero only along the boundary [4, Lemma 5.15].*

2. *We can compute the scaling of the highest order coefficients, A and D, with the system size.*

Due to Remark 20, we can estimate the EPR in a manner that does not depend on the system size by normalizing by the appropriate scaling. For instance, for the adsorption/desorption system on an $N \times N$ lattice, since the boundary scales as $O(N)$,

and because the commutator is non-zero only at the boundaries between sub-lattices, that is, $C(\sigma, \sigma^{x,y}) = 0$ if $x, y$ are not in the boundaries of different sub-lattices, the coefficient $A_{\text{Lie}}$ in (2.27) scales like $O(N)$ too. Specifically for the Lie splitting, the per-particle highest order coefficient of the RER (appearing in Equation (2.24) as "$A$") would be $A/N$. We do this for all estimates in this work, i.e., they are per-lattice-size estimates. Note that the linear scaling is a property of systems that change a single lattice site per jump, such as the adsorption/desorption example. Accordingly, other systems can have different scaling for the computation of the highest order coefficients, see for example the diffusion system in B.5.

### 2.3.1 Impact of lattice decomposition on reversibility retention

One of the choices a practitioner has to make when using parallel KMC is the decomposition of the lattice, for example checkerboard versus stripes (see Figures 2.1, 2.2). Selecting the right decomposition can affect the load-balancing of the algorithm as well as the feasibility of the run. For instance, it may be that the size of the lattice is large enough to prohibit even loading the whole system into the memory of a processor. Then, splitting the lattice into blocks, as in Figure 2.1, can often bypass this issue, whereas splitting into stripes may not be advantageous.

However, the choice of decomposition also has an effect on the error the splitting method generates per time step, both for bounded time intervals [4] and for long simulations [26]. This error is controlled by the commutator associated with the scheme, and the analysis in [4] shows that a decomposition into stripes results to reduced error due to the smaller size of the boundary region when compared to a block decomposition when blocks and stripes have the same width, see Figures 2.1 and 2.2. By approximating the EPR, we can quantify the long-time effect that the change of decomposition has to the reversibility that each scheme retains per time step. To discuss those issues, we simulated an adsorption/desorption process and

used the samples to estimate the EPR. For details about the setup of the example see B.4, information about the estimators is in B.3.

In Figure 2.3 we can see how *sensitive* each scheme is to different decompositions of the lattice. In both cases, the schemes have a smaller EPR estimate when using a stripe versus a block decomposition (where the width of the blocks matches the width of the stripes, see Figures 2.1, 2.2). In fact, the Strang scheme has consistently better performance in controlling the loss of reversibility with respect to $\Delta t$.

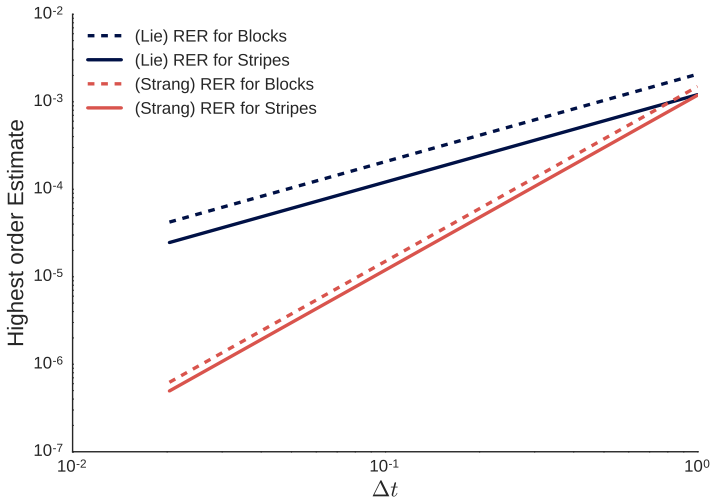


**Figure 2.4.** Approximations to the RER of the form $A \cdot \Delta t^{p-1}$ for the same adsorption/desorption system as with Figure 2.3 for the Lie splitting and Strang splitting. Lie looks more sensitive to changes in the decomposition of the lattice.

## 2.4 Derivations and General Theory

In this section, we present the general theory concerning the asymptotic behavior of the entropy production rate (EPR) of a scheme $Q_{\Delta t}$. The arguments presented here, although mirroring some of the ideas from our previous work [26], also take into account the additional discrepancy term, $I(Q_{\Delta t}|P_{\Delta t})$. Although we handle only the case that $L$ is split into $L_1 + L_2$, the arguments can also generalize to splittings with

more components, e.g. $L_1 + L_2 + L_3$. In fact, the arguments can readily generalize to schemes that are not splittings, as long as there is an expression for the error like the one in Lemma 18. Nevertheless, we will continue to consider splitting methods in this section.

**Remark 21.** *An implicit assumption in the parallel schemes used in Section 2.3 was that the splitting of the generator $L$ into $L_1+L_2$ was such that if $q(\sigma,\sigma')=0$ for some pair of states $(\sigma,\sigma')$, then $q_1(\sigma,\sigma')=q_2(\sigma,\sigma')=0$. This is imposed by the domain decomposition of the lattice and we also assume this throughout for any splitting of $L$, although the methodology can be extended to other splittings too.*

### 2.4.1 Decomposition of the Entropy Production Rate

To better understand the Entropy Production Rate, we shall first decompose it into two pieces, the relative entropy rate, Equation (2.22), and a "discrepancy" term (Equation (2.23)) that we will denote with $I$.

**Theorem 22.** *Let $\Delta t > 0$ and $P_{\Delta t}$ be a transition probability, with stationary distribution $\mu$, that satisfies detailed balance. Then, if $Q_{\Delta t}$ is an approximation coming from a numerical scheme, we have that*

$$\mathrm{EPR}(Q_{\Delta t}) = H(Q_{\Delta t}|P_{\Delta t}) + I(Q_{\Delta t}|P_{\Delta t}). \tag{2.29}$$

*Proof.* In Equation (2.16), we defined *entropy production rate* corresponding to $Q_{\Delta t}$ as

$$\mathrm{EPR}(Q_{\Delta t}) = \frac{1}{\Delta t}\sum_{\sigma,\sigma'} \mu_{\Delta t}(\sigma)Q_{\Delta t}(\sigma,\sigma')\log\left(\frac{Q_{\Delta t}(\sigma,\sigma')}{Q_{\Delta t}(\sigma',\sigma)}\right), \tag{2.30}$$

We will first introduce the reversible $P_{\Delta t}$ in Equation (2.30) as

$$\Delta t\cdot\mathrm{EPR}(Q_{\Delta t}) = \sum_{\sigma,\sigma',\sigma'\neq\sigma} \mu_{\Delta t}(\sigma)Q_{\Delta t}(\sigma,\sigma')\log\left(\frac{Q_{\Delta t}(\sigma,\sigma')P_{\Delta t}(\sigma,\sigma')P_{\Delta t}(\sigma',\sigma)}{P_{\Delta t}(\sigma,\sigma')P_{\Delta t}(\sigma',\sigma)Q_{\Delta t}(\sigma',\sigma)}\right).$$

This allows us to split the logarithm into three pieces.

$$\begin{aligned}\Delta t \cdot \mathrm{EPR}(Q_{\Delta t}) = &\sum_{\sigma,\sigma'} \mu_{\Delta t}(\sigma) Q_{\Delta t}(\sigma,\sigma') \log\left(\frac{Q_{\Delta t}(\sigma,\sigma')}{P_{\Delta t}(\sigma,\sigma')}\right) \\ &+ \sum_{\sigma,\sigma'} \mu_{\Delta t}(\sigma) Q_{\Delta t}(\sigma,\sigma') \log\left(\frac{P_{\Delta t}(\sigma,\sigma')}{P_{\Delta t}(\sigma',\sigma)}\right) \\ &+ \sum_{\sigma,\sigma'} \mu_{\Delta t}(\sigma) Q_{\Delta t}(\sigma,\sigma') \log\left(\frac{P_{\Delta t}(\sigma',\sigma)}{Q_{\Delta t}(\sigma',\sigma)}\right).\end{aligned} \tag{2.31}$$

We shall now show that the middle sum is equal to zero. By our assumptions, we know that the pair $(P_{\Delta t}, \mu)$ satisfies detailed balance, i.e. $\mu(\sigma')/\mu(\sigma) = P_{\Delta t}(\sigma,\sigma')/P_{\Delta t}(\sigma',\sigma)$. Therefore,

$$\sum_{\sigma,\sigma'} \mu_{\Delta t}(\sigma) Q_{\Delta t}(\sigma,\sigma') \log\left(\frac{P_{\Delta t}(\sigma,\sigma')}{P_{\Delta t}(\sigma',\sigma)}\right) = \sum_{\sigma,\sigma'} \mu_{\Delta t}(\sigma) Q_{\Delta t}(\sigma,\sigma') \left[\log(\mu(\sigma')) - \log(\mu(\sigma))\right]. \tag{2.32}$$

Looking at each sum in Equation (2.32) separately and using that $\mu_{\Delta t}(\sigma') = \sum_{\sigma} \mu_{\Delta t}(\sigma) Q_{\Delta t}(\sigma,\sigma')$, we have

$$\begin{aligned}\sum_{\sigma}\sum_{\sigma'} \mu_{\Delta t}(\sigma) Q_{\Delta t}(\sigma,\sigma') \log(\mu(\sigma')) &= \sum_{\sigma'} \mu_{\Delta t}(\sigma') \log(\mu(\sigma')), \\ \sum_{\sigma}\sum_{\sigma'} \mu_{\Delta t}(\sigma) Q_{\Delta t}(\sigma,\sigma') \log(\mu(\sigma)) &= \sum_{\sigma} \mu_{\Delta t}(\sigma) \log(\mu(\sigma)).\end{aligned}$$

Thus, the right-hand side of Equation (2.32) is equal to zero and we have,

$$\begin{aligned}\Delta t \cdot \mathrm{EPR}(Q_{\Delta t}) = &\sum_{\sigma,\sigma'} \mu_{\Delta t}(\sigma) Q_{\Delta t}(\sigma,\sigma') \log\left(\frac{Q_{\Delta t}(\sigma,\sigma')}{P_{\Delta t}(\sigma,\sigma')}\right) \\ &+ \sum_{\sigma,\sigma'} \mu_{\Delta t}(\sigma) Q_{\Delta t}(\sigma,\sigma') \log\left(\frac{P_{\Delta t}(\sigma',\sigma)}{Q_{\Delta t}(\sigma',\sigma)}\right).\end{aligned}$$

or

$$\mathrm{EPR}(Q_{\Delta t}) = (H(Q_{\Delta t}|P_{\Delta t}) + I(Q_{\Delta t}|P_{\Delta t}))/\Delta t.$$

□

Note that, even though the EPR and the RER are always non-negative, the discrepancy, $I$, is not. If $Q_{\Delta t}$ is reversible, then $\mathrm{EPR}(Q_{\Delta t}) = 0 \Rightarrow H(Q_{\Delta t}|P_{\Delta t}) = -I(Q_{\Delta t}|P_{\Delta t})$. If, in addition, $Q_{\Delta t} \neq P_{\Delta t}$, then the RER is positive, which implies that $I$ would be negative.

### 2.4.2 Asymptotic Behavior of Entropy Production Rate

In Theorem 22, we saw that we can express the entropy production rate (EPR) of a scheme as a sum of two different components, the relative entropy rate (Equation (2.22)) and the discrepancy (Equation (2.23)). The objective of this section is the study of each component separately via asymptotic expansions with respect to $\Delta t$. Then, at the end of the section we have an asymptotic result for the EPR based on the individual results and Equation (2.21).

In the derivations that follow, we will often refer to the distances between different states of the state space. A path $\vec{z}$ of length $|\vec{z}| = n$ between states $\sigma, \sigma'$ corresponds to a sequence $\vec{z} = (z_0, \ldots, z_n)$, with $z_0 = \sigma, z_n = \sigma'$, and distinct intermediate states $z_i$ such that $\prod_{i=0}^{n} q(z_i, z_{i+1}) > 0$, $q$ being the transition rates of the CTMC of interest. The set of all paths between those two states will be denoted by $\mathrm{Path}(\sigma \to \sigma')$. We can thus define the distance between two states with respect to a fixed CTMC by the length of the smallest path, $d(\sigma, \sigma')$. More formally,

$$d(\sigma, \sigma') := \begin{cases} \min\{|\vec{z}| : \vec{z} \in \mathrm{Path}(\sigma \to \sigma')\}, & \mathrm{Path}(\sigma \to \sigma') \neq \emptyset, \\ \infty, & \mathrm{Path}(\sigma \to \sigma') = \emptyset. \end{cases} \tag{2.33}$$

The function $d$ is the geodesic distance and is always calculated with respect to the transition rates $q$ of the exact process, $P_{\Delta t}$. In the time-reversible case, it is simple to show that $d$ is actually a metric of the state space, as it is symmetric and satisfies the triangle inequality. We also define the *diameter* with respect to $d$ as $\mathrm{diam}(S) = \max_{(\sigma,\sigma')\in S\times S}\{d(\sigma,\sigma')\}$.

We introduced the use of the geodesic distance (2.33) in Section 8 of [26]. For schemes that satisfy the requirement in Remark 21, the addition of this graph-theoretic perspective can both simplify and generalize the computations. For completeness, we include the result concerning the long-time behavior of the scheme with respect to the RER [26, Theorem 8.6].

**Lemma 23.** *Let $P_{\Delta t}(\sigma,\sigma') = e^{L\Delta t}\delta_{\sigma'}(\sigma)$ and $Q_{\Delta t}(\sigma,\sigma')$ be an approximation of $P_{\Delta t}$ based on an operator splitting scheme and $\mu_{\Delta t}$ the stationary measure corresponding to $Q_{\Delta t}$. Then, if the scheme is of order $p$, $\mathrm{diam}(S) \geq p$, and $C(\sigma,\sigma') \neq 0$ for at least one pair $\sigma,\sigma' \in S$ such that $d(\sigma,\sigma') = p$, we have that*

$$H(Q_{\Delta t}|P_{\Delta t}) = O(\Delta t^{p-1}),$$

*for $\Delta t \leq 1$.*

Note that the assumption $\mathrm{diam}(S) \geq p$ is not particularly restrictive for the original Markov process. For example, in lattice systems with adsorption/desorption, diffusion, or other spin-flip mechanisms, consider states that require three jumps of the original Markov process to go from one state to the other. Then $\mathrm{diam}(S) \geq 3$, which is sufficient for the schemes considered here, as the maximum order of the local error is attained by the Strang splitting and is equal to three. Also, checking the existence of a pair $(\sigma,\sigma')$ for which the commutator $C$ is not zero is just a matter of computation.

**Lemma 24.** *Under the assumptions of Lemma 23, the discrepancy has the same order with the RER. That is,*

$$I(Q_{\Delta t}|P_{\Delta t}) = O(\Delta t^{p-1}).$$

*Proof.* To show this, we expand $I$ in an asymptotic expansion around $\Delta t$. We demonstrate that the coefficient of the $\Delta t^{p-1}$ term comes from considering the states $\sigma, \sigma'$ such that $d(\sigma, \sigma') \leq p$ and that the dominant order is indeed equal to $p-1$ for small $\Delta t$. We note here that the assumptions on the order, $p$, and the commutator from Lemma 23 are the only assumptions on the operator splitting scheme.

We defind the discrepancy term in Equation (2.23) as:

$$\Delta t \cdot I(Q_{\Delta t}|P_{\Delta t}) = \sum_{\sigma,\sigma'} \mu_{\Delta t}(\sigma) Q_{\Delta t}(\sigma, \sigma') \log\left(\frac{P_{\Delta t}(\sigma', \sigma)}{Q_{\Delta t}(\sigma', \sigma)}\right).$$

Using the atanh representation of the logarithm [26, Equation 5.8] and its expansion, we get that

$$\Delta t \cdot I(Q_{\Delta t}|P_{\Delta t}) = \sum_{\sigma,\sigma'} \mu_{\Delta t}(\sigma) Q_{\Delta t}(\sigma, \sigma') \cdot 2 \sum_{k=0}^{\infty} \frac{1}{2k+1} \left(\frac{P_{\Delta t}(\sigma', \sigma) - Q_{\Delta t}(\sigma', \sigma)}{Q_{\Delta t}(\sigma', \sigma) + P_{\Delta t}(\sigma', \sigma)}\right)^{2k+1}. \tag{2.34}$$

In the proof of Lemma 23 (Theorem 5.2 in [26]), we use our knowledge of the asymptotic behavior of $P_{\Delta t} \pm Q_{\Delta t}$ for small $\Delta t$ [26, Equations (5.3), (5.4)] to infer the behavior of ratios of those quantities. That is,

$$\frac{P_{\Delta t}(\sigma', \sigma) - Q_{\Delta t}(\sigma', \sigma)}{P_{\Delta t}(\sigma', \sigma) + Q_{\Delta t}(\sigma', \sigma)} = \frac{C(\sigma', \sigma)}{2Q_{\Delta t}(\sigma', \sigma) + C(\sigma', \sigma)\Delta t^p} \Delta t^p + o(\Delta t^p). \tag{2.35}$$

We assume that all $\sigma, \sigma'$ satisfy $d(\sigma, \sigma') = p$, i.e. they are $p$ jumps apart. For notational brevity, we define

$$M(\sigma, \sigma') := \frac{C(\sigma, \sigma')}{C(\sigma, \sigma') + 2L_Q^p(\sigma, \sigma')/p!}, \tag{2.36}$$

where $L_Q^p$ represents all the terms in the expansion of $Q_{\Delta t}$ that are of order $p$ (see Equation (B.19) in appendix). Then, for $k > 0$, we have that

$$\begin{aligned}
&\sum_{\sigma,\sigma'} \mu_{\Delta t}(\sigma) Q_{\Delta t}(\sigma, \sigma') \cdot 2 \sum_{k=1}^{\infty} \frac{1}{2k+1} \left( \frac{P_{\Delta t}(\sigma', \sigma) - Q_{\Delta t}(\sigma', \sigma)}{P_{\Delta t}(\sigma', \sigma) + Q_{\Delta t}(\sigma', \sigma)} \right)^{2k+1} \\
=&\sum_{\sigma,\sigma'} \mu_{\Delta t}(\sigma) L_Q^p(\sigma, \sigma') \frac{2\Delta t^p}{p!} \left( \operatorname{atanh}(M(\sigma', \sigma)) - M(\sigma', \sigma) \right) + o(\Delta t^p).
\end{aligned} \tag{2.37}$$

Before we continue with the analysis of Equation (2.37), we look at the term from Equation (2.34) corresponding to the first term of the series, i.e. $k = 0$. Using Equation (2.35), we get

$$2 \sum_{\sigma,\sigma'} \mu_{\Delta t}(\sigma) Q_{\Delta t}(\sigma, \sigma') \cdot \frac{C(\sigma', \sigma)}{2Q_{\Delta t}(\sigma', \sigma) + C(\sigma', \sigma)\Delta t^p} \Delta t^p + o(\Delta t^p). \tag{2.38}$$

We notice that to get terms of order $\Delta t^p$ from the sum (2.38), we need the order of $Q_{\Delta t}(\sigma, \sigma')$ to be the same as that of $Q_{\Delta t}(\sigma', \sigma)$. We remind here that the order of the local error is equal to $p$ and that $L^i$ is the resulting operator after $i$ compositions of the generator $L$ of the original process. Therefore, if $i < p$, the ratio

$$\frac{Q_{\Delta t}(\sigma, \sigma')}{Q_{\Delta t}(\sigma', \sigma)} = \frac{L^i(\sigma, \sigma')\Delta t^i + O(\Delta t^{i+1})}{L^i(\sigma', \sigma)\Delta t^i + O(\Delta t^{i+1})} = \frac{L^i(\sigma, \sigma')}{L^i(\sigma', \sigma)} + o(\Delta t) \tag{2.39}$$

is well defined as long as $L^i(\sigma, \sigma') \neq 0$, and that is true because $d(\sigma, \sigma') = i$ implies that $L^i(\sigma, \sigma') > 0$ (see Lemma 39 in B ) and $L^j(\sigma, \sigma') = 0$ for $j < i$ [26, Lemma 8.3]. Therefore, the right-hand side of Equation (2.39) is well-defined for all $\sigma, \sigma'$ such that

$d(\sigma, \sigma') = i$, $i < p$. This finalizes the analysis of the first term of the asymptotic series for $I$, with

$$\begin{aligned} & 2\sum_{\sigma,\sigma'} \mu_{\Delta t}(\sigma) Q_{\Delta t}(\sigma,\sigma') \cdot \frac{C(\sigma',\sigma)}{2Q_{\Delta t}(\sigma',\sigma) + C(\sigma',\sigma)\Delta t^p} \Delta t^p + o(\Delta t^p) \\ & = \sum_{\sigma} \mu_Q(\sigma) \sum_{i=0}^{p-1} \sum_{\sigma' \in S_i(\sigma)} \frac{L^i(\sigma,\sigma')}{L^i(\sigma',\sigma)} C(\sigma',\sigma) \\ & \quad + \sum_{\sigma' \in S_p(\sigma)} L_Q^p(\sigma,\sigma') \frac{2}{p!} M(\sigma',\sigma) \Delta t^p + o(\Delta t^p). \end{aligned} \tag{2.40}$$

Above we use the notation $S_i(\sigma) = \{\sigma' : d(\sigma,\sigma') = i\}$. Now, if we add Equations (2.37) and (2.40), the terms that involve $M(\sigma',\sigma)$ cancel. Thus, we get the following asymptotic expansion for $I$.

$$\begin{aligned} I(Q_{\Delta t}|P_{\Delta t}) = & \sum_{\sigma} \mu_Q(\sigma) \sum_{i=0}^{p-1} \sum_{\sigma' \in S_i(\sigma)} \frac{L^i(\sigma,\sigma')}{L^i(\sigma',\sigma)} C(\sigma',\sigma) \Delta t^{p-1} \\ & + \sum_{\sigma,\sigma' \in S_p(\sigma)} \mu_{\Delta t}(\sigma) L_Q^p(\sigma,\sigma') \frac{2}{p!} \operatorname{atanh}(M(\sigma',\sigma)) \Delta t^{p-1} \\ & + o(\Delta t^{p-1}). \end{aligned} \tag{2.41}$$

□

Equation (2.41) is the basis for our estimation of $I$ for small $\Delta t$, which is used in Section 2.3. An immediate implication of Theorem 22 and Lemmas 23 and 24 is the next result, which provides the scaling of the EPR with respect to $\Delta t$.

**Theorem 25.** *Let $\Delta t \in (0,1)$. Let $P_{\Delta t}(\sigma,\sigma') = e^{L\Delta t}\delta_{\sigma'}(\sigma)$ and $Q_{\Delta t}(\sigma,\sigma')$ be an approximation of $P_{\Delta t}$ based on a splitting scheme and $\mu_{\Delta t}$ the stationary measure corresponding to $Q_{\Delta t}$. In addition, let $P_{\Delta t}$ satisfy detailed balance and* $\operatorname{diam}(S) \geq p$. *Then,*

$$\mathrm{EPR}(Q_{\Delta t}) = O(\Delta t^{p-1}). \tag{2.42}$$

Finally, note that $\mathrm{EPR}(Q_{\Delta t}) = O(\Delta t^{p-1})$ implies that the corresponding RER has order $O(\Delta t^{p-1})$, or better. In other words, a numerical scheme of high leading order in EPR is also more accurate in sampling from the stationary regime [26].

## 2.5 Conclusions

We introduced the entropy production rate (EPR) as a means to quantify the loss of reversibility for operator splitting schemes applied to Parallel Kinetic Monte Carlo. We showed estimation of the EPR does not require the knowledge of the stationary distribution and depends on the transition probabilities of the scheme. Since the transition probabilities for stochastic particle systems are usually not available, or difficult to explicitly compute, we derived *a posteriori* estimators of the EPR and connected the parameters of the scheme with a quantitative assessment of the loss of reversibility. We demonstrated this fact with an application to lattice KMC with adsorption/desorption dynamics, which we simulated using SPPARKS [56], and a comparison between two splitting schemes, Lie and Strang. Theory and simulations show that the Strang splitting retains more reversibility per time step compared to Lie and is more stable with respect to changes in the decomposition of the lattice (blocks versus stripes, see Figure 2.3).

The proposed framework for Parallel KMC, can be applied to more than computational schedule comparison. In essence, the EPR can be used as an information criterion that allows practitioners to judge the fine details of the scheme itself, like the time step and which lattice decompositions retain more reversibility (see Figure 2.3). The EPR can also be used as a diagnostic observable to assess the reversibility of the scheme used by simulating a system of smaller size than the one of interest. In this way, issues with the scheme can be discovered early on using a much smaller system for diagnostics, different schemes can be compared, and parameters tuned to minimize the loss of reversibility.

Though we only considered operator splitting schemes in the context of parallel lattice KMC, the idea of using the EPR for the quantification of the loss of reversibility can be used on other schemes too, as long as an expression for their local error exists and is computable. For instance, an extension of this work can be used to quantify the loss of reversibility for schemes used for thermostated Molecular Dynamics simulations [44], for example for Langevin dynamics [45].

# CHAPTER 3

# QUANTIFYING MODEL BIAS WITH CONCENTRATION INEQUALITIES

We illustrate how we can combine the sharpness of the goal-oriented divergence with a collection of concentration inequalities to construct sharp bounds for model bias. The new divergences allow us to trade available information about a partially-known model $P$ for guarantees on model bias that are applicable to a whole class of quantities of interest (QoIs). We show how the bounds behave on a series of simple examples.

## 3.1 Goal-Oriented Divergence

In this section, we introduce the goal-oriented divergence (GO), which allows us to get sharp bounds on the bias without the issues that other information metrics suffer from (see [37]). The GO divergence was first defined in [21], following the work in [16].

Consider two distributions $P, Q$, with $Q$ absolutely continuous with respect to $P$, and a QoI $f : \mathcal{X} \to \mathbb{R}$ such that $\log \mathbb{E}_P[e^{cf}]$ is finite in a neighborhood of the origin. Then, we will call the quantities $\Xi_{\pm}(Q\|P; f)$ the "*goal-oriented* (GO) *divergences*", defined as:

$$\Xi_{\pm}(Q\|P; f) := \inf_{c>0} \left\{ \frac{1}{c} \log M_P(\pm c; \tilde{f}) + \frac{1}{c} R(Q\|P) \right\} \tag{3.1}$$

$$M_P(c; f) = \mathbb{E}_P[e^{cf}], \tag{3.2}$$

$$\tilde{f} = f - \mathbb{E}_P[f]. \tag{3.3}$$

Because of its dependence on the KL and the moment-generating function (MGF), the GO divergence has the following properties:

1. $\Xi_{\pm}(Q\|P; f) \geq 0$ for all $Q \ll P$ and QoIs $f$.

2. $\Xi_{\pm}(Q\|P; f) = 0$ if and only if $Q = P$ or $f$ is constant P-almost surely.

3. Linearization:

$$\Xi_{\pm}(Q\|P; f) = \pm\sqrt{\mathrm{var}_P[f]}\sqrt{2R(Q\|P)} + O(R(Q\|P))$$

We remind that $Q \ll P$ means that $Q$ is absolutely continuous with respect to $P$: If $P(A) = 0$ for some $P$-measurable set $A$, then $Q(A) = 0$.

Our primary interest in the GO divergences stems from the fact that they form bounds for the model bias:

$$-\Xi_{-}(Q\|P; f) \leq \mathbb{E}_Q[f] - \mathbb{E}_P[f] \leq \Xi_{+}(Q\|P; f). \tag{3.4}$$

We will refer to the bounds in (3.4) as the GO bounds.

We will not reproduce the derivation of $\Xi_{\pm}(Q\|P; f)$ and the proof of (3.4) here; see [21] and [16] for details. For the convenience of the reader we provide a sketch of why (3.4) is true. By the variational principle for the relative entropy (see [23]), we know that for any bounded function $f$ we have

$$\log M_P(c; f) = \sup_{Q:\ Q \ll P} \{\mathbb{E}_Q[f] - R(Q\|P)\}$$

and thus for any distribution $Q$ such that $Q \ll P$:

$$\mathbb{E}_Q[f] \leq \log M_P(c; f) + R(Q \mid\mid P).$$

Replacing $f$ by $c(f - \mathbb{E}_P[f])$ with $c > 0$ and optimizing over $c$ yields the upper bound from (3.4). The lower bound is derived in a similar manner.

Finally, note that Property 3 implies that when $R(Q\|P)$ is small, then

$$|\mathbb{E}_Q[f] - \mathbb{E}_P[f]| \leq \sqrt{\mathrm{var}_P[f]}\sqrt{2R(Q\|P)} + O(R(Q\|P)). \tag{3.5}$$

The GO divergences provides sharp bounds for model bias and remains discriminating in the presence of large data [37]. However, its dependence on the MGF, $\mathbb{E}_P[e^{c\tilde{f}}]$, makes its practical use challenging, especially when $P$ is partially known, as is the case in Bayesian inference or applications that require a coarse-graining method. In addition, estimators of the MGF can suffer from exponentially-increasing variance as $c$ is increasing, requiring extensive sampling and/or the use of multi-level Monte Carlo methods; see [47]. To bypass the computation of the MGF, we will instead leverage the information we have about $P$ and the QoI $f$ to construct bounds on the MGF that are valid in (at least) a neighborhood of the origin.

## 3.2 Concentration inequalities

We wish to use the GO bounds for the quantification of the bias but in order to compute them we need the CGF of $\tilde{f}$ under $P$ (see (3.1)). When only some aspects of $P$ are known, for example, if $P$ is a complicated posterior distribution that is known up to a constant, then alternatives to the GO bound that require less information would be more useful. A particular strategy is to seek a function $\Phi_P(c)$ as well as a family of QoIs, $\mathcal{F}_P$, such that:

$$M_P(c; \tilde{g}) \leq \Phi_P(c) \text{ for all } g \in \mathcal{F}_P \tag{3.6}$$

and all $c$ in a neighborhood of the origin.

One application that makes use of such bounds to the MGF is the study of *concentration* of random variables; see, for example, Chapter 2. of [62]. A classical result is the *Chernoff bound*: For any $a \in \mathbb{R}$ and $c > 0$,

$$P(X \geq a) = P(e^{cX} \geq e^{ca}) \leq \frac{\mathbb{E}[e^{cX}]}{e^{ca}}. \tag{3.7}$$

Thus, one way to understand the tail behavior is the derivation of bounds to the MGF (or, equivalently, the CGF) of $X$. Via this approach, more informative inequalities can be derived. For instance, the *tail* bound in (3.8) implies that the distribution of $X - \mu$ is dominated in the tails by a Normally-distributed random variable:

$$P(|X - \mu| \geq \sigma a) \leq 2e^{-a^2} \text{ for all } a \geq 0, \tag{3.8}$$

where $\sigma$ is the standard deviation of $X$. One can show that (3.8) is equivalent to a Gaussian bound on the MGF of $X$ [Theorem 2.1][62], i.e., there exists a $\sigma_B > 0$ such that

$$\mathbb{E}[e^{cX}] \leq e^{c^2\sigma_B^2/2} \text{ for all } c \in \mathbb{R}. \tag{3.9}$$

A random variable that satisfies (3.9) is called a *sub-Gaussian* random variable with parameter $\sigma_B$. This class of random variables is broad—it includes all bounded random variables. However, we will see that more sharp bounds than (3.9) exist and can be ordered in terms of the information that they require. Nevertheless, the discussion on bounds of the MGF motivates the derivation of more bounds for the bias, the general form of which is described in Theorem 26.

**Theorem 26.** *Consider distributions $P, Q$ such that $Q \ll P$ and the pair $(\Phi_P(c), \mathcal{F}_P)$, where $\Phi_P : \mathbb{R} \to \mathbb{R}^+$ that satisfies (3.6) for all $c$ in a neighborhood of the origin. Then,*

$$U_-(Q\|P; \mathcal{F}_P) \leq \mathbb{E}_Q[g] - \mathbb{E}_P[g] \leq U_+(Q\|P; \mathcal{F}_P) \text{ for all } g \in \mathcal{F}_P, \tag{3.10}$$

*where*

$$U_\pm(Q\|P; \mathcal{F}_P) := \inf_{c>0} \left\{ \frac{1}{c} \log \Phi_P(c) + \frac{1}{c} R(Q\|P) \right\}. \tag{3.11}$$

*Proof.* The bounds to the bias in (3.10) are a consequence of (3.4) and (3.6). □

The set $\mathcal{F}_P$ can often be described explicitly in terms of the QoIs that satisfy the requirements that $\Phi_P(c)$ imposes. For instance, for the Bennet bound (3.13) with parameters $b, \sigma_B$:

$$\mathcal{F}_P = \{g : g(x) \leq b \text{ for all } x \in \text{supp}\{P\},\ \text{var}_P[g] \leq \sigma_B^2\}. \tag{3.12}$$

Therefore, given a specific QoI $f$, we have to find an appropriate pair $(\Phi_P(c), \mathcal{F}_P)$ that will allow us to use Theorem 26 to bound the model bias with respect to $f$. As such, $f$ also has to belong in $\mathcal{F}_P$. It is more natural to separate the discussion for the cases of bounded and unbounded QoIs.

### 3.2.1 Bounded observables

Many QoIs are bounded; probabilities of events are a prime example, smooth functions of random variables with bounded support are another. Bounded random variables are sub-Gaussian [62, Chapter 2.] and sharp bounds for their MGFs can be derived, which can then be turned to bias bounds through Theorem 26. We next showcase the bounds that we will use in this work, although this list is not extensive by any means and other concentration inequalities can also be derived [57].

**Bennet bound** [20, Lemma 2.4.1]: Consider an observable function $f$ such that $f(X) \leq b$, $X \sim P$, for some $b \geq 0$. Setting $\mu := \mathbb{E}_P[f(X)], \tilde{b} := b - \mu$, we have

$$M_P(c; \tilde{f}) \leq \frac{\tilde{b}^2}{\tilde{b}^2 + \sigma_B^2} \exp(-c\sigma_B^2/\tilde{b}) + \frac{\sigma_B^2}{\tilde{b}^2 + \sigma_B^2} \exp(c\tilde{b}), \tag{3.13}$$

for all $c \geq 0$ and where $\sigma_B^2$ is any bound of $\mathrm{var}_P[f]$.

**Bennet-$(a, b)$ bound** [20, Corollary 2.4.5]: If $f$ is such that $a \leq f(X) \leq b$, $X \sim P$, then we can set $\sigma_B^2 = (\mu - a)(b - \mu)$ in the Bennet bound to get

$$M_P(c; \tilde{f}) \leq \frac{\tilde{b}}{b-a} \exp(c\hat{a}) - \frac{\hat{a}}{b-a} \exp(c\tilde{b}) \text{ for all } c \in \mathbb{R}. \tag{3.14}$$

The right-hand side of (3.14) is the MGF of a Bernoulli-distributed random variable with values $\{a, b\}$, as this is the distribution with the most "spread" around the mean value between all bounded random variables in $[a, b]$.

**Hoeffding bound** [31]: When $f$ is bounded as in the Bennet-$(a, b)$ case, we can bound Bennet-$(a, b)$ by a Gaussian MGF, giving us the Hoeffding MGF bound:

$$M_P(c; \tilde{f}) \leq \exp(c^2(b-a)^2/8) \text{ for all } c \in \mathbb{R}. \tag{3.15}$$

The Hoeffding bound is independent of the location of the mean $\mu$ within the interval $(a, b)$ and only depends on the length of the interval. As such, it requires the least amount of information about $f$ and $P$ but, as can be seen in Section 3.5, it can also fail to capture important information about $f$.

By Lemma 2.4.1 and Corollary 2.4.5 in [62], we can order the bounds in terms of accuracy, assuming that $f$ is bounded in $[a, b]$ and $\sigma_B^2 \leq (\mathbb{E}_P[f] - a)(b - \mathbb{E}_P[f])$:

$$M_P(c; \tilde{f}) \leq \text{Bennet} \leq \text{Bennet-(a,b)} \leq \text{Hoeffding}. \tag{3.16}$$

| Name | Conditions on $f, P$ |
|---|---|
| Hoeffding (3.15) | $a \leq f(X) \leq b$ |
| Bennet-$(a,b)$ (3.14) | $a \leq f(X) \leq b$ |
| Bennet (3.13) | $f(X) \leq b$, $\mathrm{var}_P[f] \leq \sigma_B^2$ |
| sub-Gaussian (3.17) | $M_P(c;\tilde{f}) \leq \exp(\sigma_B^2 c^2/2)$ for all $c \in \mathbb{R}$ |
| sub-Exponential (3.18) | $M_P(c;\tilde{f}) \leq \exp(\sigma_B^2 c^2/2)$ for all $c \in (-1/\beta, 1/\beta)$ |
| GO bound (3.1) | $\exp(c(f - \mathbb{E}_P[f]))$: well-defined in a region of 0 |

**Table 3.1.** The different MGF bounds, along with the conditions they impose on $P$ and $f$. In terms of information requirements, the Hoeffding bound requires the least amount, but it is also the least tight. As information requirements grow, $U_{\pm}(Q\|P;\mathcal{F}_P)$ approach $\Xi_{\pm}(Q\|P;f)$.

Note that the assumption on $\sigma_B^2$ is necessary for the Bennet bound to be more accurate than the Bennet-(a,b).

### 3.2.2 Unbounded observables

If the QoI is unbounded but has appropriate tail decay properties, we can still use concentration inequalities. In (3.9) we wrote one such property, the Gaussian decay of the tails, which we repeat here.

**sub-Gaussian behavior:** $X$ is a sub-Gaussian random variable if there exists a $\sigma_B > 0$ such that

$$M_P(c;\tilde{X}) \leq \exp(c^2\sigma_B^2/2) \text{ for all } c \in \mathbb{R} \tag{3.17}$$

By using the definition of the MGF, we can show that $\sigma_B^2$ is an upper bound of the variance of $X$ under $P$. However, in contrast with the Bennet inequality (see (3.13)), $\sigma_B^2$ is not just any bound of the variance and needs to be computed prior to using the bound in (3.17); as an example of what could $\sigma_B^2$ be, see McDiarmid's inequality[57, Section 2.2.3]. Apart from (3.17), there are other equivalent ways to show that $X$ is sub-Gaussian [62, Chapter 1.2]—proving the bound in (3.8) is one example.

In general, sub-Gaussianity is a strong assumption for an unbounded random variable. For example, if $X \sim P = \text{Laplace}(1)$, i.e., a two-sided exponential distribution, then $M_P(cX) = 1/(1-c^2),\ |c| < 1$, which cannot be bounded by any $\exp(c^2\sigma_B^2/2)$ for all $c$. We can relax the requirement by instead asking for a local bound of the MGF.

**sub-Exponential behavior:** $X$ is a sub-exponential random variable [62, Section 2.1.3] if there exist $\sigma_B,\ \beta$: positive such that

$$M_P(c; \tilde{X}) \leq \exp(c^2\sigma_B^2/2) \text{ for all } |c| \leq 1/\beta. \tag{3.18}$$

$\sigma_B^2$ and $\beta$ depend on the tail decay of the distribution $P$. For example, for $X \sim P = \text{Laplace}(1)$, we would have

$$M_P(cX) \leq \exp(2c^2) \text{ for } |c| < 1/2. \tag{3.19}$$

As long as the MGF of $X$ exists in a region around zero, Gaussian bounds are always valid locally [15] with $\text{var}_P X \leq \sigma_B^2$.

Unbounded random variables can have other types of tail decay. See Section 3.6 for a discussion about Poisson decay.

## 3.3 Examples

In order to understand better both the mathematical and the computational points that come up, we will use the bounds on a series of simple situations. Unless otherwise stated, the KL is not calculated for a specific pair of $Q, P$. Instead, the distribution $P$ is fixed and we consider $U_\pm(Q\|P; \mathcal{F}_P)$ for $R(Q\|P) = \eta^2$ for $\eta$ in some pre-specified range. That allows us to get a *non-parametric* picture for the worst-case bias.

## 3.4 Exponential distribution

We first consider the UQ bounds for the case of the exponential distribution as an example of a distribution with unbounded support and an MGF that is not well-behaved everywhere.

Consider $P$: exponential distribution with parameter $\lambda_P = 1$ and QoI: $f(X) = X$. The MGF of $P$ is

$$M_P(c) = \frac{1}{1-c},$$

and thus the MGF is well-defined in $(0, 1)$. Now, if $Q$ is such that $R(Q\|P) = \eta^2$ for some $\eta > 0$, then

$$\Xi_{\pm}(Q\|P; f) = \inf_{c\in(0,1)} \left\{ \frac{1}{c} M_P(\pm c; \tilde{f}) + \frac{\eta^2}{c} \right\}. \tag{3.20}$$

Because the MGF is only well-defined close to zero, we had to constrain the optimization problem in Equation (3.20) (see also the definition of $\Xi_{\pm}(Q\|P; f)$ in Equation (3.1)).

The distribution $P$ exhibits sub-exponential behavior (see Section 3.2.2), i.e., its MGF is locally bounded by a Gaussian MGF. To see this, we first constrain $c$ in the interval $(-0.5, 0.5)$. Then, by expanding the MGF:

$$M_P(c) = 1 + c + \frac{c^2}{1-c} \leq 1 + c + 2c^2 \leq \exp(c + c^2/(2\sigma_B^2)), \tag{3.21}$$

where $\sigma_B = 1/2$. If we have information on the location of $\lambda_P$, e.g., from data, then we can change the length of the interval. $(-0.5, 0.5)$ is picked here for illustration purposes.

Using the MGF bound (3.21) and the definition of $U_{\pm}(Q\|P; \mathcal{F}_P)$ from (3.11), we can get another bound on the bias. However, this bound applies equally to all QoIs in the family $\mathcal{F}_P$:

$$\mathcal{F}_P = \{g : \mathbb{R} \to \mathbb{R} : \mathbb{E}_P[g] = 1, \mathrm{var}_P[g] \leq 1/4\}. \tag{3.22}$$

Additionally, $U_{\pm}(Q\|P; \mathcal{F}_P)$ require less information from $P$ compared to the GO bound (3.20). Figure 3.1 has a comparison of the two bounds along with the bias for the case that $Q$ is also an exponential distribution with $\lambda_Q \in (1.01, 10)$.

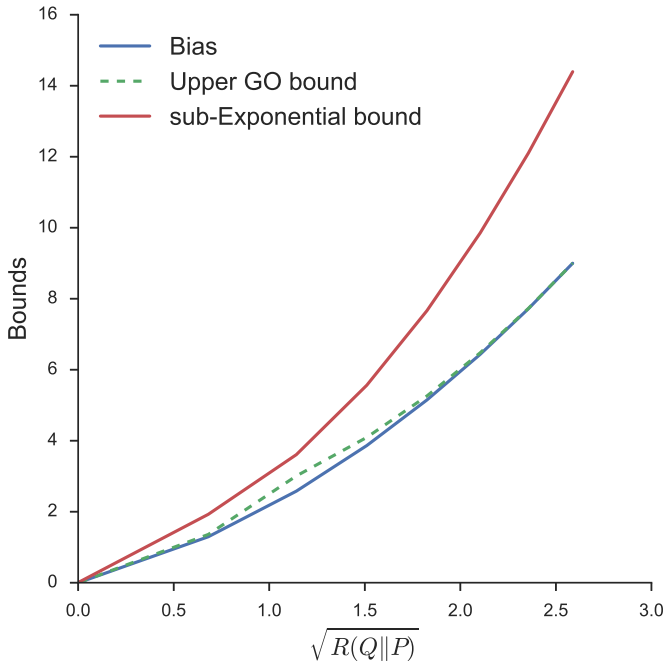


**Figure 3.1.** Comparison of the two bounds with the bias $1 - \lambda_Q, \lambda_Q \in (1.01, 10)$. The sub-exponential bound is considerably less sharp as the KL increases as it paints a broad picture of worst-case performance over the family of observables $\mathcal{F}_P$ from (3.22).

**Remark 27.** *The bias is an unbounded function of the KL divergence in this example—a consequence of the observable $f(X) = X$ being unbounded under $P$. Therefore, any decrease in KL divergence translates to an improvement in worst-case bias, in sharp contrast with the truncated Normal example (see Section 3.5) where small improvements to large values of the KL may not help much in reducing uncertainty.*

## 3.5 Truncated Normal

Suppose $Y \sim N(\mu, \sigma^2)$ and $a, b$ such that $a < b$. Then, the random variable $X = Y | a \leq Y \leq b$ follows the truncated Normal (TN) distribution in $[a, b]$. We will use the TN as an example of a distribution with bounded support.

Let $P$ be a TN distribution in $[-1, 1]$ with $\mu = 0.7$ and $\sigma^2 = 0.5$. Then, with $U_{\pm}(Q\|P; \mathcal{F}_P)$ we can compute robust bounds to the bias if the approximation, $Q$, is such that $R(Q\|P) = \eta^2$, for $\eta^2 \in [0.01, 1]$. Note that the bounds in Figure 3.2 provide a non-parametric result; the only thing we assume about the distribution $Q$ is that $R(Q\|P) = \eta^2$.

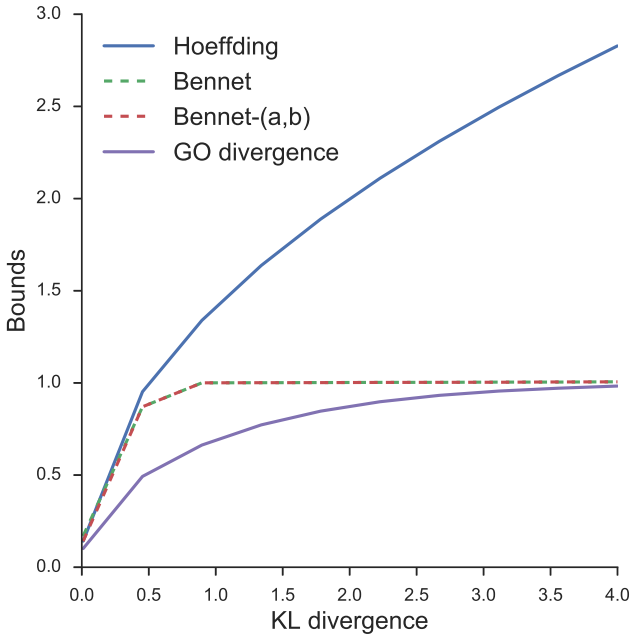


**Figure 3.2.** Comparison of the different bounds for the bias in the truncated Normal example (see Section 3.5), assuming that the QoI is $f(X) = X$. This plot makes no assumptions on the form of $Q$ except that $R(Q\|P) = \eta^2 \in (0.0, 4.0)$. Notice that Bennet and Bennet-$(a, b)$ correctly capture the bound of the GO divergence for large values of the KL whereas the Hoeffding is only accurate for small values of the KL, i.e., at the linearized regime of the GO bounds. Only the upper bounds for the bias are being shown here.

A particular feature of Figure 3.2 is that for bounded QoIs we may have less to gain in terms of worst-case bias by improving a large value of the KL divergence. For the bounds that are tight, this is an indication of when simple improvements can offer any advantage in reducing uncertainty versus having to use additional resources, e.g., sampling additional data, including more complicated parameterized families, etc. However many bounds can exist and not all of them will be tight, for example, the Hoeffding bound in Figure 3.2.

Even for the tightest $U$-like bounds, i.e., Bennet and Bennet-$(a, b)$, there is some discrepancy with the GO bound. This is to be expected; whereas the GO bound is applied for a specific observable, the $U$ bounds are sharp over a whole class of QoIs. We give up the need for most of the information on the QoI of $f$ to move from the GO to the $U$ bounds — from UQ for $f$ to UQ for the class of observables $\mathcal{F}_P$. We trade sharpness for generality and applicability.

## 3.6 Poisson tails

The concentration properties of an observable can differ substantially from the quadratic decay of the tails of the Normal distribution that we have seen so far. For instance, quantities of interest that have appropriate properties, e.g., self-boundedness [12], also have Poisson concentration. That is, for those observables $f$, we have

$$\log M_P(c; \tilde{f}) \leq \exp(c) - c - 1. \tag{3.23}$$

Then the bound (3.23) can be used with $U_{\pm}(Q\|P; \mathcal{F}_P)$ to derive the corresponding bound on the bias.

The study of Poisson tail behavior is as extensive as in the Gaussian case; see [57, Section 3.3.5] for more references and [11] for conditions that imply Poisson-like behavior.

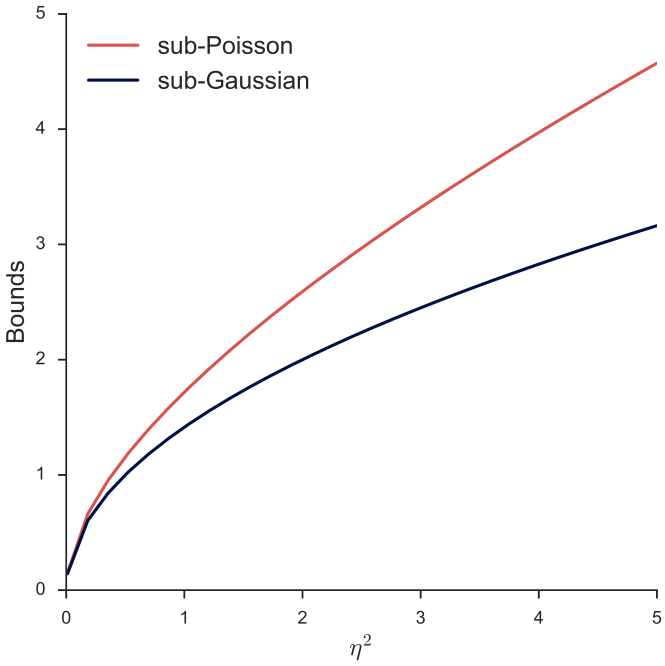


**Figure 3.3.** Growth of $U(\eta^2; \mathcal{F})$ with the sub-Poisson bound from Inequality (3.23) versus the sub-Gaussian bound $\exp(c^2/2)$ (see Inequality (3.17)) and $\eta^2 \in [0.01, 5]$.

## 3.7 Discussion

We have introduced concentration bounds as a systematic way to trade information about $P$ and the QoI $f$ with guarantees for the model bias that are applicable over a whole family of QoIs $\mathcal{F}_P$. Such bounds are tractable to compute while still retaining the properties of the GO divergence. In follow-up work, we will apply them to a variety of examples, such as model misspecification, failure probability calculations, variational inference, etc.

We have not explored data assimilation for the bounds in this thesis, however this would be a necessary part for a practical implementation. As can be seen in Table 3.1, the Bennet bound requires the expected value of the QoI as well as a bound for its variance under $P$. Given a finite sampling budget from $P$, those quantities can be estimated up to a certain precision. In the I.I.D. case, we can exactly quantify the

required precision in terms of estimator variance via the central limit theorem. We explore this matter more in follow-up work.

Another interesting point concerns extending the tools presented here to the case of Markov processes and path-space QoIs. We discussed an example of interest in the first two chapters where we considered approximations to an exact process $P$ by another process $Q$ (arising from the parallel scheme) and discussed how to discriminate between different schemes via the path-space relative entropy and the relative entropy rate. In [21], a version of the GO divergences was also derived that covers such a case and bounds the bias of QoIs with respect to path distributions. An extension of the ideas of this chapter for such situations would be useful and is in our future plans.

Finally, throughout the chapter we assumed that the KL divergence is a known quantity. In situations where $Q$ can be easily sampled (and $P$ is known up to a multiplicative constant), we may employ thermodynamic integration [9, 46] to estimate the KL. Upper bounds to the KL may also be usable in some cases [37]. Ultimately, the accuracy of the bounds discussed here depends crucially on knowledge of the KL by other methods.

# APPENDIX A

# SUPPLEMENTARY MATERIAL FOR CHAPTER 1

## A.1 Coefficients of the relative entropy rate for Lie and Strang

For the adsorption-desorption example considered in section 1.5 of the main text we need to estimate the highest-order coefficients $A, B$ for Lie and Strang, respectively. To accomplish this, we have to collect all the coefficients of $\Delta t$ and $\Delta t^2$ that appear in the expansion of RER in the proof of Theorem 7. The result is a summable series for each coefficient. For Lie, we have

$$A = \mathbb{E}_{\mu_L(\sigma)}\left[\sum_{x,y\in\Lambda} F_L(\sigma,\sigma^{x,y})\right] = \sum_{\sigma}\mu_L(\sigma)\sum_{x,y\in\Lambda} F_L(\sigma,\sigma^{x,y}), \tag{A.1}$$

$$F_L(\sigma,\sigma') := C_L(\sigma,\sigma')M_L(\sigma,\sigma') - 2L_L^2[\delta_{\sigma'}](\sigma)(\operatorname{arctanh}(M_L(\sigma,\sigma')) - M_L(\sigma,\sigma')), \tag{A.2}$$

$$M_L(\sigma,\sigma') := C_L(\sigma,\sigma')/(L_L^2[\delta_{\sigma'}(\sigma)] + C_L(\sigma,\sigma')),$$

where we remind the reader that $L_L^2$ stands for all the coefficients of $\Delta t^2/2$ in the expansion of the Lie splitting and $C_L(\sigma,\sigma') = 1/2[L_1, L_2][\delta_{\sigma'}](\sigma)$ is the Lie commutator term. Similarly, for the Strang case,

$$B = \mathbb{E}_{\mu_S(\sigma)}\left[\sum_{x,y,z\in\Lambda} F_S(\sigma,\sigma^{x,y,z})\right] = \sum_{\sigma}\mu_S(\sigma)\sum_{x,y,z\in\Lambda} F_S(\sigma,\sigma^{x,y,z}), \tag{A.3}$$

$$F_S(\sigma,\sigma') := C_S(\sigma,\sigma')M_S(\sigma,\sigma') - 2L_S^3[\delta_{\sigma'}](\sigma)(\operatorname{arctanh}(M_S(\sigma,\sigma')) - M_S(\sigma,\sigma')), \tag{A.4}$$

$$M_S(\sigma,\sigma') := C_S(\sigma,\sigma')/(L_S^3[\delta_{\sigma'}](\sigma) + C_S(\sigma,\sigma')). \tag{A.5}$$

Since both (A.1) and (A.3) are expected values, we can estimate them as ergodic averages.

## A.2 Proofs for the supporting results

### A.2.1 Representation of semigroups

One of the main tools that we used in the main text was that we could represent the transition probabilities of an operator splitting scheme via a convergent power series.

**Lemma 28.** *Let $L$ be linear & bounded operator, $L : C_b(S) \to C_b(S)$, $L = L_1 + L_2$ with $L_1, L_2$ also linear & bounded. Then, let $Q_{\Delta t}$ be an approximation to $P_{\Delta t}$ via a splitting method. That is*

$$Q_{\Delta t}(\sigma, \sigma') = \prod_{i=1}^{n} \exp(a_i \Delta t L_1) \exp(b_i \Delta t L_2) \delta_{\sigma'}[\sigma], \tag{A.6}$$

*for $n \in \mathbb{N}$ and $a_i, b_i \in \mathbb{R}, i = 1, \ldots, n$ and $n \in \mathbb{N}$. Given this, we have,*

$$Q_{\Delta t}(\sigma, \sigma') = \sum_{k=0}^{\infty} \frac{\Delta t^k}{k!} L_Q^k[\delta_{\sigma'}](\sigma), \ \sigma, \sigma' \in S. \tag{A.7}$$

*Proof.* From the boundedness of the operators $L_1, L_2$, we can represent the corresponding semigroups as power series. Thus, for any constants $a_1, b_1 \in \mathbb{R}$, we have that

$$\exp(a_1\Delta t L_1)\exp(b_1\Delta t L_2) = \sum_{k=0}^{\infty}\frac{a_1^k\cdot\Delta t^k}{k!}L_1^k\sum_{m=0}^{\infty}\frac{b_1^m\Delta t^m}{m!}L_2^m.$$

This a classical Cauchy product of convergent series and thus is equal to

$$\exp(a_1\Delta t L_1)\exp(b_1\Delta t L_2) = \sum_{k=0}^{\infty}\left(k!\cdot\sum_{m=0}^{k}\frac{a_1^m L_1^m}{m!}\cdot\frac{b_1^{k-m}L_2^{k-m}}{(k-m)!}\right)\cdot\frac{\Delta t^k}{k!}. \tag{A.8}$$

Then, if $Q_{\Delta t}(\sigma,\sigma') = \exp(a_1\Delta t L_1)\exp(b_1\Delta t L_2)\delta_{\sigma'}(\sigma), \sigma,\sigma'\in S$, we could write $Q_{\Delta t}$ as the expansion in the lemma by having

$$L_Q^k := k!\cdot\sum_{m=0}^{k}\frac{a_1^m L_1^m}{m!}\cdot\frac{b_1^{k-m}L_2^{k-m}}{(k-m)!}. \tag{A.9}$$

Now, let us assume that we have $a_i, b_i\in\mathbb{R}, i = 0,\ldots,n$, for some $n$ positive integer. As mentioned in the statement of the lemma, this gives us a representation of the splitting in the general form

$$Q_{\Delta t}(x,y) = \prod_{i=1}^{n}\exp(a_i\Delta t L_1)\exp(b_i\Delta t L_2)\delta_y[x]. \tag{A.10}$$

Since all semigroups have well-defined series expansions, we can apply the idea in (A.8) iteratively to all intermediate expansions until $Q_{\Delta t}$ is represented by a power series in terms of $\Delta t$. The resulting coefficients of $\Delta t^k/k!$ at the end of this process are denoted by $L_Q^k$ (in the spirit of (A.9)). □

Using the representation of the operator splitting scheme, we can also show the existence of the commutator, which captures the local error of the scheme.

**Lemma 29** (Local order of error & commutator)**.** *Let* $P_{\Delta t}(\sigma, \sigma') = e^{L\Delta t}\delta_{\sigma'}(\sigma)$ *and* $Q_{\Delta t}(\sigma, \sigma')$ *an approximation of* $P_{\Delta t}$ *via a splitting scheme. Then, there is a function* $C : S \times S \to \mathbb{R}$ *and an integer* $p$, $p > 1$, *such that*

$$P_{\Delta t}(\sigma, \sigma') = Q_{\Delta t}(\sigma, \sigma') + C(\sigma, \sigma')\Delta t^p + o(\Delta t^p). \tag{A.11}$$

*Proof.* Let $\sigma', \sigma \in S$. Then, we have,

$$P_{\Delta t}(\sigma, \sigma') - Q^{\text{Lie}}_{\Delta t}(\sigma, \sigma') = (e^{L\Delta t} - e^{L_1\Delta t}e^{L_2\Delta t})\delta_{\sigma'}(\sigma). \tag{A.12}$$

and we would like to show that

$$P_{\Delta t}(\sigma, \sigma') - Q^{\text{Lie}}_{\Delta t}(\sigma, \sigma') = \frac{1}{2}[L_1, L_2]\delta_{\sigma'}(\sigma)\Delta t^2 + O(\Delta t^3). \tag{A.13}$$

We assume that $L, L_1, L_2$ are all bounded operators. Given this, Lemma 28 applies for the semigroups $e^{L_1\Delta t}$ and $e^{L_2\Delta t}$. Thus it is valid to express all semigroups in (A.12) as series expansions. Then,

$$(e^{L\Delta t} - e^{L_1\Delta t}e^{L_2\Delta t})\delta_{\sigma'}(\sigma) = \left(\sum_{k=0}^{\infty}\frac{\Delta t^k}{k!}L^k - \sum_{m=0}^{\infty}\frac{\Delta t^m}{m!}L_1^k \cdot \sum_{n=0}^{\infty}\frac{\Delta t^n}{n!}L_2^n\right)\delta_{\sigma'}(\sigma).$$

Lemma 28 applies for the product of semigroups $e^{L_1\Delta t}e^{L_2\Delta t}$, which leads to a power series representation for it,

$$e^{L_1\Delta t}e^{L_2\Delta t} = I + (L_1 + L_2)\Delta t + \left(L_1^2 + L_2^2 + 2L_1L_2\right)\frac{\Delta t^2}{2} + O(\Delta t^3). \tag{A.14}$$

The corresponding expansion for the exact semigroup is

$$e^{L\Delta t} = I + (L_1 + L_2)\Delta t + (L_1^2 + L_2^2 + L_1L_2 + L_2L_1)\frac{\Delta t^2}{2} + O(\Delta t^3). \tag{A.15}$$

Subtracting (A.14) from (A.15) and applying on $\delta_{\sigma'}(\sigma)$ gives the result. □

### A.2.2 Order of the Relative Entropy Rate

To motivate the normalization of the relative entropy rate (RER) by $\Delta t$ in the main text, we showed that for two different transition probabilities $Q^A_{\Delta t}, Q^B_{\Delta t}$, the RER has at least an order of $\Delta t$.

**Lemma 30.** *Let $L_A, L_B$ be bounded generators of Markov Processes, $L_A \neq L_B$, with corresponding transition probabilities $Q^A_{\Delta t}, Q^B_{\Delta t}$. Then,*

$$H(Q^B_{\Delta t}|Q^A_{\Delta t}) = O(\Delta t).$$

*Proof.* We note first that the corresponding transition probabilities are

$$Q^A_{\Delta t}(\sigma, \sigma') = \exp(\Delta t L_A)\delta_{\sigma'}(\sigma),\ Q^B_{\Delta t}(\sigma, \sigma') = \exp(\Delta t L_B)\delta_{\sigma'}(\sigma).$$

Then, if $\mu$ is the stationary distribution associated with $Q^B_{\Delta t}$, we have

$$H(Q^B_{\Delta t}|Q^A_{\Delta t}) = \Delta t \sum_{(\sigma,\sigma')\in I} \mu(\sigma)\frac{(q_A(\sigma, \sigma') - q_B(\sigma, \sigma'))^2}{q_A(\sigma, \sigma') + q_B(\sigma, \sigma')} + O(\Delta t), \tag{A.16}$$

where $I = \{(\sigma, \sigma') \in S^2 : q_A(\sigma, \sigma') + q_B(\sigma, \sigma') > 0\}$. To derive this, we work similarly as in the case of the Theorems in the main text, using the representations of $Q^A_{\Delta t}, Q^B_{\Delta t}$ via expansions and subsequently expanding the logarithm in the definition of the relative entropy rate. As $Q^B_{\Delta t} \neq Q^A_{\Delta t}$, only the first term of their expansions match, which then gives Equation (A.16) as a result. □

### A.2.3 Generators and Graph Distance

In Section 8 of the main text, we discussed a more general perspective based on graph theory, that allowed us to compute the order of the relative entropy rate based

on properties of both the schemes (order of the local error) and the original system. For the convenience of the reader, we include the relevant definition of the distance between states.

**Definition 31** (Distance between states)**.** *Let $q$ be the transition rates of a Continuous Time Markov Process over a countable state space $S$. Then, let $\sigma, \sigma' \in S$, $\sigma \neq \sigma'$. The distance $d_q$ between the two states is defined as*

$$d_q(\sigma, \sigma') := \min \{|\vec{z}| : \vec{z} \in \mathrm{Path}(\sigma \to \sigma')\} \tag{A.17}$$

*In the case that the two states are disconnected, i.e.* $\mathrm{Path}(\sigma \to \sigma') = \emptyset$*, then* $d(\sigma, \sigma') = +\infty$*. Given those distances, one can also define the diameter of the space as*

$$\mathrm{diam}(S) = \max_{(\sigma,\sigma') \in S \times S} \{d(\sigma, \sigma')\}.$$

**Lemma 32.** *Let $L$ be an infinitesimal generator defined as,*

$$L[f](\sigma) = \sum_z q(\sigma, z)(f(z) - f(\sigma))$$

*for $q$ transition rates. Let any $n \in \mathbb{N}$ and let $\sigma'$ be a fixed state, we have*

$$\{\sigma : L^n[\delta_{\sigma'}](\sigma) \neq 0\} \subseteq \{\sigma : d(\sigma, \sigma') \leq n\} = B_n(\sigma').$$

*Proof.* We fix a $\sigma' \in S$. Then, if $L[\delta_{\sigma'}](\sigma) = q(\sigma, \sigma') \neq 0$, by the definition of the distance $d$, $\sigma$ belongs in the set $\{\sigma : d(\sigma, \sigma') \leq 1\}$. We proceed by induction, assuming our proposition as a fact for $n = k \in \mathbb{N}$. The implication is that

$$\{\sigma : L^k \delta_{\sigma'}(\sigma) \neq 0\} \subseteq \{\sigma : d(\sigma, \sigma) \leq k\}.$$

Now, we can prove it for $n = k + 1$. Let $\sigma \in \{z : d(z, \sigma') > k + 1\}$. Then,

$$\begin{aligned} L^{k+1}[\delta_{\sigma'}](\sigma) &= L[L^k[\delta_{\sigma'}](\sigma)] \\ &= \sum_z q(\sigma, z) \left( L^k[\delta_{\sigma'}](z) - L^k[\delta_\sigma](\sigma) \right) \end{aligned} \tag{A.18}$$

But then, since $d(\sigma, \sigma') > k+1 \Rightarrow d(\sigma, \sigma') > k$ and thus $L^k[\delta_{\sigma'}](\sigma) = 0$. Thus, (A.18) becomes,

$$L^{k+1}[\delta_{\sigma'}](\sigma) = \sum_z q(\sigma, z) L^k[\delta_{\sigma'}](z) \tag{A.19}$$

Note that sum (A.19) is over all states $z$ such that $d(\sigma, z) = 1$ since otherwise $q(\sigma, z) = 0$. But then, if $d(\sigma, \sigma') > k + 1$, it must be the case that $d(z, \sigma') > k$ and thus, from our assumption, we get that $L^k[\delta_{\sigma'}](z) = 0$ for all such $z$ states. Thus, we have that:

$$\text{If } d(\sigma, \sigma') > k + 1 \Rightarrow L^{k+1}[\delta_{\sigma'}](\sigma) = 0.$$

This is the result we wanted. □

To study general operator splitting schemes, we defined restrictions of the generator $L$ given a subset $A$ of $S \times S$.

**Definition 33** (Restriction of a generator)**.** *Let us have set $A$ with $A \subset S \times S$ and $L$ be an infinitesimal generator of a Markov process with associated transition rates $q$. Then, the restriction $L|_A$ of $L$ is defined as*

$$L|_A[f](\sigma) = \sum_{\sigma' \in S} q_A(\sigma, \sigma') \left( f(\sigma') - f(\sigma) \right), \ \sigma \in S, \tag{A.20}$$

*where $q_A(\sigma, \sigma') = q(\sigma, \sigma') \cdot \chi_A(\sigma, \sigma')$, $\chi_A$ is the characteristic function of set $A$ and $f$ is continuous and bounded function on the state space $S$.*

**Lemma 34.** *Let us have the state space $S$ and $S \times S = A \cup B, A \cap B = \emptyset$, along with generators $L_1 = L|_A$, $L_2 = L|_B$. We fix $\sigma' \in S$ and $k, m \in \mathbb{N}$. Then,*

$$\left\{\sigma : L_1^k \left[L_2^m[\delta_{\sigma'}]\right](\sigma) \neq 0\right\} \subseteq \{\sigma : d(\sigma, \sigma') \leq k + m\}.$$

*Proof.* Let $k, m \in \mathbb{N}$ and $\sigma' \in S$. Then, we first note that $L_1^k = (L|_A)^k = (L^k)|_A$ (and similarly for $L_2^m$). This is a consequence of the way the restriction of $L$ is defined in (A.20). Since $L_1^k$ is a restriction of $L^k$, we have that if $L^k[\delta_{\sigma'}](\sigma) = 0$ then $L_1^k[\delta_{\sigma'}](\sigma) = 0$. This can be expressed as

$$\{\sigma : L_1^k[\delta_{\sigma'}](\sigma) \neq 0\} \subseteq \{\sigma : L^k[\delta_{\sigma'}](\sigma) \neq 0\}.$$

Now, we can use Lemma 32 to get

$$\{\sigma : L_1^k[\delta_{\sigma'}](\sigma) \neq 0\} \subseteq \{\sigma : d(\sigma, \sigma') \leq k\}. \tag{A.21}$$

Given this, we can address compositions of the operators $L_1, L_2$. The proof for the case $L_1^k[L_2^m[\delta_{\sigma'}]](\sigma)$ is similar to the one for Lemma 32. For the sake of the reader, we provide the argument for the case of $k = 1$ and the rest can be done by induction. Let $\sigma, \sigma'$ states such that $d(\sigma, \sigma') > m + 1$. Then we have

$$L_1[L_2^m[\delta_{\sigma'}]](\sigma) = \sum_z q_1(\sigma, z) \left(L_2^m[\delta_{\sigma'}](z) - L_2^m[\delta_{\sigma'}](\sigma)\right) \tag{A.22}$$

In the sum of (A.22), we only consider states $z$ such that $q_1(\sigma, z) \neq 0$. Let $z \in A$ be such a state. Since $d(\sigma, \sigma') > m + 1$ and $L_2$ is a restriction of $L$, $L_2^m[\delta_{\sigma'}](\sigma) = 0$ (from relation (A.21)). Because $q_1(\sigma, z) = q(\sigma, z) \neq 0$, it holds that $d(\sigma, z) \leq 1$. Thus, applying the triangle inequality gives that $d(z, \sigma') \geq m$ for all such $z$ and so $L_2^m[\delta_{\sigma'}](z) = 0$. As a result, if $\sigma \notin B_{m+1}(\sigma')$, we get that $L_1[L_2^m[\delta_{\sigma'}]](\sigma) = 0$. □

The idea in Lemma 34 can also be applied to more complicated compositions of $L_1, L_2$ in the same way.

## A.3 RER for a fully connected system

Systems such that the transitions rate $q(\sigma, \sigma')$ are positive for all states $\sigma, \sigma'$, will be characterized as *fully connected*. In those, we can have a transition from some state to any other state in one step. As we will see, this simplifies the asymptotic work on the RER. This case is considered only for its simplicity, as systems of the above type are not easily parallelized.

In the proof that follows, we will need to quantify the dependence of various relations involving $P_{\Delta t}$ and $Q_{\Delta t}$ to $\Delta t$.

$$P_{\Delta t}(\sigma, \sigma') - Q_{\Delta t}(\sigma, \sigma') = C(\sigma, \sigma')\Delta t^p + o(\Delta t^p), \tag{A.23}$$

$$P_{\Delta t}(\sigma, \sigma') + Q_{\Delta t}(\sigma, \sigma') = 2\delta_{\sigma'}(\sigma) + 2q(\sigma, \sigma')\Delta t + o(\Delta t) \tag{A.24}$$

$$= 2Q_{\Delta t}(\sigma, \sigma') + C(\sigma, \sigma')\Delta t^p + o(\Delta t^p). \tag{A.25}$$

Those formulas are based on Lemma 2.2 in the main text. Thus, we can write a theorem for the RER and its relation to the various quantities discussed for the case of a fully connected system.

**Theorem 35.** *Let $P_{\Delta t}(\sigma, \sigma') = e^{L\Delta t}\delta_{\sigma'}(\sigma)$ and $Q_{\Delta t}(\sigma, \sigma')$ be an approximation of $P_{\Delta t}$ based on a splitting scheme. Also, let $\mu_{\Delta t}$ the invariant measure associated with $Q_{\Delta t}$. Then, if the order of the local error between $Q_{\Delta t}$ and $P_{\Delta t}$ is equal to $p, p > 1$, and if the system is fully connected,*

$$H(Q_{\Delta t}|P_{\Delta t}) = \sum_{\sigma \in S} \mu_{\Delta t}(\sigma) \sum_{\substack{\sigma' \in S \\ \sigma' \neq \sigma}} \frac{(C(\sigma, \sigma'))^2}{2q(\sigma, \sigma')} \Delta t^{2p-2} + o(\Delta t^{2p-2}). \tag{A.26}$$

We remind that $C$ stands for the commutator associated to the scheme, as defined in Lemma 2.2 of the main text.

*Proof.* Given $x > 0$ and by the definition of $tanh^{-1}$,

$$\log(x) = 2\text{atanh}\left(\frac{x-1}{x+1}\right) = 2\sum_{k=0}^{\infty}\frac{1}{2k+1}\left(\frac{x-1}{x+1}\right)^{2k+1}. \tag{A.27}$$

This expansion of the logarithm converges for every $x > 0$, as can be seen by applying the root convergence test. Next, we expand the logarithm in the definition of the RER according to (A.27) in order to get

$$\begin{aligned}\Delta t \cdot H(Q_{\Delta t}|P_{\Delta t}) &= 2\sum_{\sigma,\sigma'}\mu_Q(\sigma)Q_{\Delta t}(\sigma,\sigma')\frac{Q_{\Delta t}(\sigma,\sigma') - P_{\Delta t}(\sigma,\sigma')}{Q_{\Delta t}(\sigma,\sigma') + P_{\Delta t}(\sigma,\sigma')} + F(\Delta t),\\ F(\Delta t) &:= 2\sum_{\sigma,\sigma'}\mu_Q(\sigma)Q_{\Delta t}(\sigma,\sigma')\sum_{k=1}^{\infty}\frac{1}{2k+1}\left(\frac{Q_{\Delta t}(\sigma,\sigma') - P_{\Delta t}(\sigma,\sigma')}{Q_{\Delta t}(\sigma,\sigma') + P_{\Delta t}(\sigma,\sigma')}\right)^{2k+1}.\end{aligned} \tag{A.28}$$

Note that $p > 1$ implies a match of the terms in the expansion of $P_{\Delta t}$ and $Q_{\Delta t}$ corresponding to $\Delta t^0$ and $\Delta t^1$. To progress, we first need asymptotic results for the ratios $((P_{\Delta t} - Q_{\Delta t})/(Q_{\Delta t} + P_{\Delta t}))^k$, which are proved in Lemma 36. Thus, we get that $F(\Delta t) = O(\Delta t^{3p-2})$ and we have,

$$\begin{aligned}\Delta t \cdot H(Q_{\Delta t}|P_{\Delta t}) = -2\sum_{\substack{\sigma,\sigma'\\ \sigma\neq\sigma'}}\mu_{\Delta t}(\sigma)Q_{\Delta t}(\sigma,\sigma')\frac{P_{\Delta t}(\sigma,\sigma') - Q_{\Delta t}(\sigma,\sigma')}{Q_{\Delta t}(\sigma,\sigma') + P_{\Delta t}(\sigma,\sigma')}\\ + O(\Delta t^{3p-2}).\end{aligned} \tag{A.29}$$

Now we use Equation (A.25) on the denominator of the fraction appearing in (A.29). This allows us to approximate the fraction in (A.29) and, after simplifications, we get

$$\Delta t \cdot H(Q_{\Delta t}|P_{\Delta t}) = -\sum_{\sigma,\sigma'} \mu_{\Delta t}(\sigma)\left(P_{\Delta t}(\sigma,\sigma') - Q_{\Delta t}(\sigma,\sigma')\right) + \sum_{\sigma,\sigma'} \mu_{\Delta t}(\sigma) G(\Delta t;\sigma,\sigma')$$
$$= \sum_{\sigma,\sigma'} \mu_{\Delta t}(\sigma) G(\Delta t;\sigma,\sigma'), \tag{A.30}$$

since $\sum_{\sigma'} P_{\Delta t}(\sigma,\sigma') - Q_{\Delta t}(\sigma,\sigma') = 0$ for any $\sigma \in S$. Thus next we will quantify the dependence of $G(\Delta t;\sigma,\sigma')$ in $\Delta t$. Let us fix $\sigma,\sigma' \in S$ and neglect the sum. Then we have

$$G(\Delta t;\sigma,\sigma') = \frac{-\left(P_{\Delta t}(\sigma,\sigma') - Q_{\Delta t}(\sigma,\sigma')\right)\Delta t^p C(\sigma,\sigma')}{2Q_{\Delta t}(\sigma,\sigma') + \Delta t^p C(\sigma,\sigma') + o(\Delta t^p)} + o(\Delta t^{2p}). \tag{A.31}$$

We can apply equation (A.23) to the numerator of (A.31) to further expose the $\Delta t$. This manipulation leads to,

$$-\left(P_{\Delta t}(\sigma,\sigma') - Q_{\Delta t}(\sigma,\sigma')\right)\Delta t^p C(\sigma,\sigma') = -(C(\sigma,\sigma'))^2 \Delta t^{2p} + o(\Delta t^{2p}).$$

Now, we need to also take care of the dependence of $\Delta t$ in the denominator. To do that, we can separate the fraction to an approximation plus remaining terms and use that $Q_{\Delta t}(\sigma,\sigma') = q(\sigma,\sigma')\Delta t + o(\Delta t)$ to get

$$\frac{-(C(\sigma,\sigma'))^2\Delta t^{2p} + o(\Delta t^{2p})}{2Q_{\Delta t}(\sigma,\sigma') + \Delta t^p C(\sigma,\sigma') + o(\Delta t^p)} = -\frac{(C(\sigma,\sigma'))^2}{2q(\sigma,\sigma')}\Delta t^{2p-1} + O(\Delta t^{2p}). \tag{A.32}$$

Therefore $G(\Delta t;\sigma,\sigma') = O(\Delta t^{2p-1})$, when $\sigma \neq \sigma'$. We should also tend to the case that $\sigma' = \sigma$, in which $Q_{\Delta t}(\sigma,\sigma') = 1 + O(\Delta t)$ and so

$$-\frac{(C(\sigma,\sigma))^2}{2Q_{\Delta t}(\sigma,\sigma)}\Delta t^{2p} = -\frac{(C(\sigma,\sigma))^2}{2}\Delta t^{2p} + O(\Delta t^{2p+1}). \tag{A.33}$$

It follows that $G(\Delta t;\sigma,\sigma) = O(\Delta t^{2p})$. So, combining (A.32) and (A.30), we get the result. □

Note that, compared to the first theorem appearing in the main text, Theorem 35 shows a substantial difference in the order of the RER for the two splittings. What it suggests is that the RER, apart from depending on the way $L$ is decomposed into $L_1, L_2$, is also influenced by the properties of the original process. Finally, note the assumption we made on the commutator. As is hinted in the proof, violation of this assumption would mean that $H(Q^{\text{Lie}}_{\Delta t}|P_{\Delta t}) = O(\Delta t^2)$ instead of $O(\Delta t^1)$, which matches the order in Theorem 35. We comment on this point in Section 8 of the main text. We can see that the commutator and the order of the local error both make an appearance in the $\Delta t$-expansion of the RER. This implies that criteria for splitting selection, which were applicable for short time intervals, remain relevant for long-time errors as well.

For the convenience of the reader, we include Lemma 36, versions of which were also used in the main text.

**Lemma 36.** *Let $P_{\Delta t}(\sigma, \sigma') = e^{L\Delta t}\delta_{\sigma'}(\sigma)$ and $Q_{\Delta t}$ an approximation based on a splitting scheme with order of local error $p, p > 1$. Then, given $k \in \mathbb{N}$, and $\sigma, \sigma'$ states such that $q(\sigma, \sigma') > 0, \sigma \neq \sigma'$, we have*

$$Q(\sigma, \sigma') \cdot \left( \frac{Q_{\Delta t}(\sigma, \sigma') - P_{\Delta t}(\sigma, \sigma')}{Q_{\Delta t}(\sigma, \sigma') + P_{\Delta t}(\sigma, \sigma')} \right)^k = O\left(\Delta t^{kp-(k-1)}\right). \tag{A.34}$$

*Proof.* Since $\sigma \neq \sigma'$ and $q(\sigma, \sigma') > 0$, Equation (A.24) is

$$P_{\Delta t}(\sigma, \sigma') + Q_{\Delta t}(\sigma, \sigma') = 2q(\sigma, \sigma')\Delta t + o(\Delta t). \tag{A.35}$$

Thus, using Equations (A.35) and (A.23), we get

$$\left( \frac{Q_{\Delta t}(\sigma, \sigma') - P_{\Delta t}(\sigma, \sigma')}{Q_{\Delta t}(\sigma, \sigma') + P_{\Delta t}(\sigma, \sigma')} \right)^k = \left( \frac{C(\sigma, \sigma')}{2q(\sigma, \sigma')} \right)^k \Delta t^{kp-k} + o(\Delta t^{kp-k}).$$

Since $Q_{\Delta t}(\sigma, \sigma') = q(\sigma, \sigma')\Delta t + o(\Delta t)$, we get the result. □

From Theorem 35, we can compute the first terms of the asymptotic expansion of the RER.

**Corollary 1.** *Let $P_{\Delta t}(\sigma,\sigma') = e^{L\Delta t}\delta_{\sigma'}(\sigma)$, $Q^{\mathrm{Lie}}_{\Delta t}$ be the Lie approximation and $Q^{\mathrm{Strang}}_{\Delta t}$ be the Strang approximation. Let us also denote by $\mu_L$ ($\mu_S$) the associated stationary measure to $Q^{\mathrm{Lie}}_{\Delta t}$ ($Q^{\mathrm{Strang}}_{\Delta t}$). Then, assuming that the system is fully connected,*

$$\begin{aligned}
H(Q^{\mathrm{Lie}}_{\Delta t}|P_{\Delta t}) &= \sum_{\sigma\in S}\mu_L(\sigma)\sum_{\sigma'\in S}\frac{([L_1,L_2]\delta_{\sigma'}(\sigma)))^2}{4q(\sigma,\sigma')}\Delta t^2 + O(\Delta t^3),\\
H(Q^{\mathrm{Strang}}_{\Delta t}|P_{\Delta t}) &= \sum_{\sigma\in S}\mu_S(\sigma)\sum_{\sigma'\in S}\frac{\left(1/24\left([L_1,[L_1,L_2]]-2[L_2,[L_2,L_1]]\right)\delta_{\sigma'}(\sigma)\right)^2}{2q(\sigma,\sigma')}\Delta t^4\\
&\quad + O(\Delta t^5).
\end{aligned}$$

In both cases above, the coefficients of $\Delta t^2$ and $\Delta t^4$ are expected values of an appropriate observable that includes the commutator over the corresponding stationary measure. This implies that, apart from the theoretical results we can infer from the formulas, those are quantities that can be estimated during a parallel KMC simulation as ergodic averages. More generally, in Theorem 35 we stated that

$$H(Q_{\Delta t}|P_{\Delta t}) = \mathbb{E}_{\mu_{\Delta t}}\left[\sum_{\sigma'\in S}\frac{(C(\cdot,\sigma'))^2}{2q(\cdot,\sigma')}\right]\Delta t^{2p-2} + o(\Delta t^{2p-2}). \tag{A.36}$$

So, we can estimate the top order coefficient by simulating $\sigma_{i\Delta t}\sim\mu_Q$. Applying the ergodic theorem for Markov chains,

$$\mathbb{E}_{\mu_{\Delta t}}\left[\sum_{\sigma'\in S}\frac{(C(\cdot,\sigma'))^2}{2q(\cdot,\sigma')}\right] = \lim_{T\to\infty}\frac{1}{T}\sum_{i=1}^{N_T}\Delta t\sum_{\sigma'\in S}\frac{(C(\sigma_{i\Delta t},\sigma'))^2}{2q(\sigma_{i\Delta t},\sigma')}, \tag{A.37}$$

where $N_T = [T/\Delta t]$. Therefore, given a sufficiently long path of the chain $\sigma_{i\Delta t}$ and since the commutator is a computable object for any pair of states, we can estimate the coefficient to any precision during a simulation. The coefficients for lower-order terms can also be estimated in the same fashion.

## A.4 RER-inspired higher order splitting schemes

One of the assumptions in the theorems of the main text has been the existence of a pair of states with a certain geodesic distance from each other. A choice of $L_1, L_2$ such that the commutator is zero for all such states would raise the order of the scheme used. We demonstrate this idea next in the context of the simple Markov Chain example we included in the main text (Section 8). In that example, we defined the transition rate matrix of a simple $3 \times 3$ Markov chain:

$$Q = \begin{pmatrix} -3 & 1 & 2 \\ 3 & -4 & 1 \\ 1 & 0 & -1 \end{pmatrix}.$$

Can we design a splitting of $Q$ into $A, B$ such that $H(Q^{\text{Lie}}_{\Delta t}|P_{\Delta t}) = O(\Delta t^2)$, even though the diameter of the system is equal to two? In order to accomplish this, we follow the idea illustrated in the proof of Theorem 8.6 in the main text and construct an $A$ such that the commutator $C(3,2) = [A,B]_{32} = 0$. Since $B = Q - A$, $[A,B] = [A, Q-A] = [A,Q]$, so we just need to pick a valid transition rate matrix $A$ such that $[A,Q]_{32} = 0$. That is,

$$[A,Q]_{32} = \sum_{i=1}^{3} A_{3i}Q_{i2} - Q_{3i}A_{i2} = 0. \tag{A.38}$$

A simple solution for the Equations in (A.38) is $A_{3i} = A_{i2} = 0$ for all $i$. There are now an infinite number of choices for $A$, since for any positive $a$, the matrix

$$A = \begin{pmatrix} -a & 0 & a \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \tag{A.39}$$

is a valid transition rate matrix that satisfies all of our requirements. Picking $a = 1$, computing the RER with this splitting and expanding it around zero we get,

$$H(Q^{\text{Lie}}_{\Delta t}|P_{\Delta t}) \simeq 0.223\Delta t^2 + 0.857\Delta t^3 + O(\Delta t^4) \tag{A.40}$$

and indeed the order of the $\Delta t$-normalized RER has risen to two.

## A.5 Formula for the commutator in the Lie case

We provide the full formula for the commutator associated to the Lie splitting. Let $\sigma(x) \in \{0, 1\}$ be the order parameter for all $x$ in the lattice $\Lambda$. Then, the Lie commutator is

$$\begin{aligned} C(\sigma, \sigma') = \frac{1}{2} \sum_{x,y\in\Lambda} & \left[q_1(\sigma, \sigma^x)q_2(\sigma^x, \sigma^{x,y}) - q_1(\sigma^y, \sigma^{y,x})q_2(\sigma, \sigma^y)\right] \delta_{\sigma'}(\sigma^{x,y}) \\ - \sum_{x,y\in\Lambda} & q_1(\sigma, \sigma^x) \left[q_2(\sigma^x, \sigma^{x,y}) - q_2(\sigma, \sigma^y)\right] \delta_{\sigma'}(\sigma^x) \\ - \sum_{x,y\in\Lambda} & q_2(\sigma, \sigma^y) \left[q_1(\sigma, \sigma^x) - q_1(\sigma^y, \sigma^{y,x})\right] \delta_{\sigma'}(\sigma^y). \end{aligned} \tag{A.41}$$

Note that each of the sums in (A.41) can be simplified further due to the locality of the transitions. By the definition of $q_1, q_2$ in the main text, we have

$$q_i(\sigma, \sigma^x) := q(\sigma, \sigma^x)\chi_{G_i}(x), i = 1, 2 \tag{A.42}$$

where $G_1 \cap G_2 = \emptyset$, $G_1, G_2$ being subsets of the lattice. Thus, when calculating

$$q_1(\sigma, \sigma^x)q_2(\sigma^x, \sigma^{x,y}) - q_1(\sigma^y, \sigma^{y,x})q_2(\sigma, \sigma^y) \tag{A.43}$$

we can see that (A.43) is trivially zero if $x \in G_2$ or $y \in G_1$. Also, if $x \in G_1, y \in G_2$ but $x$ is not in a lattice neighborhood of $y$, then $q_2(\sigma^x, \sigma^{x,y}) = q_2(\sigma, \sigma^y)$ and $q_1(\sigma^y, \sigma^{y,x}) =$

$q_1(\sigma, \sigma^x)$. That is, transitions that involve $x$ do not depend on $y$ and vice versa, when $x$ and $y$ are not in each other's neighborhood. So, in this case too, (A.43) is zero. Therefore, we can see that the first sum in (A.41) is much smaller, as it only involves the boundary lattice sites between the groups $G_1, G_2$. By the same argument, we can see that this is the case for the two next sums in (A.41). As a result, the Lie commutator is an object that only needs to be calculated in the boundary elements between the two lattice groups and this is a much smaller set ($O(N)$, if $N^2$ is the number of lattice sites), compared to the full lattice. For example, in Figure A.1 and for the Lie commutator, we have to sum over the elements in the green area that corresponds to each sub-lattice, counting each pair of neighbors $x, y, x \in G_1$ and $y \in G_2$, only once. The situation is similar for the Strang commutator.

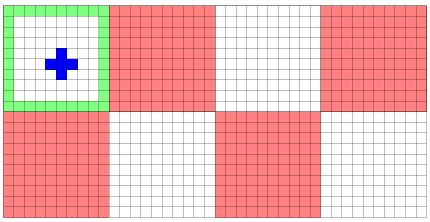

**Figure A.1.** A checkerboard decomposition of a 2D lattice. Red sub-lattices correspond to group $G_1$ and white ones to $G_2$. For comparison, a nearest neighborhood region (n.n. region) is also shown (solid black cross). Transitions involving the center of that region only depend on the state of its nearest neighbors. So, if we pick the sub-lattices much larger than the size of an n.n. region, transitions in different sub-lattices belonging to the same group are independent. A site $x$ is said to belong to the boundary of its sub-lattice if part of its n.n. region is outside that sub-lattice (the green region is the collection of all such points for the first sub-lattice). If a transition occurs at such a site $x$, then an update needs to be made to the boundary information of all other sub-lattices for which $x$ belongs to a n.n. region.

*Derivation of Equation* (A.41). Let $\sigma, \sigma'$ be states. From the text, we know that in the Lie case the commutator is equal to

$$C(\sigma, \sigma') = [L_1, L_2]/2\delta_{\sigma'}(\sigma) = \left(L_1 L_2 \delta_{\sigma'}(\sigma) - L_2 L_1 \delta_{\sigma'}(\sigma)\right)/2. \tag{A.44}$$

To progress, we have to first calculate $L_1L_2\delta_{\sigma'}(\sigma), L_2L_1\delta_{\sigma'}(\sigma)$. Due to the symmetry between the two expressions, we only need to calculate $L_1L_2\delta_{\sigma'}(\sigma)$. Thus, we have,

$$L_1[L_2\delta_{\sigma'}](\sigma) = \sum_{z\in S} q_1(\sigma, z)\left(L_2[\delta_{\sigma'}](z) - L_2[\delta_{\sigma'}](\sigma)\right). \tag{A.45}$$

First, we remind that given state $\sigma$ and lattice site $x$ then by $\sigma^x$ we denote the state such that if $y$ is another lattice site, then

$$\sigma^x(y) = \begin{cases} \sigma(y), & y \neq x, \\ 1-\sigma(y), & y = x. \end{cases} \tag{A.46}$$

Next, we will assume that the transition rates $q$ (and thus $q_1, q_2$) have the following property.

$$q(\sigma, \sigma') = \begin{cases} q(\sigma, \sigma^x) > 0, & \sigma' = \sigma^x, \\ 0, & \text{else}, \end{cases} \tag{A.47}$$

Thus, by (A.47), Equation (A.45) can be re-written as

$$L_1[L_2\delta_{\sigma'}](\sigma) = \sum_{x\in\Lambda} q_1(\sigma, \sigma^x)\left(L_2[\delta_{\sigma'}](\sigma^x) - L_2[\delta_{\sigma'}](\sigma)\right). \tag{A.48}$$

Using the definition of $L_2$ and assumption in (A.47), we have

$$\begin{aligned} L_1L_2\delta_{\sigma'}(\sigma) = &\sum_{x,y\in\Lambda} q_1(\sigma, \sigma^x)q_2(\sigma^x, \sigma^{x,y})\left(\delta_{\sigma'}(\sigma^{x,y}) - \delta_{\sigma'}(\sigma^x)\right) \\ &- \sum_{x,y\in\Lambda} q_1(\sigma, \sigma^x)q_2(\sigma, \sigma^y)\left(\delta_{\sigma'}(\sigma^y) - \delta_{\sigma'}(\sigma)\right) \end{aligned} \tag{A.49}$$

To get to $L_2L_1$ from $L_1L_2$, we can use the symmetry of the formulas and just switch rates $q_1$ with $q_2$ in (A.49). Then, subtracting the two gives the result. □

## A.6 Further details on asynchronous algorithm for parallel KMC

For completeness, we include a step-by-step description of the algorithm for parallel lattice KMC. The background section in the paper has more details and references.

Assuming that we can decompose the lattice $\Lambda$ into sub-lattices $\Lambda_i, i = 1, \ldots, n$ so that transitions in some sub-lattices are independent from transitions in others, like in Figure A.1. Let $G_2$ be the set of red sub-lattices and $G_1$ the set of white sub-lattices. Then, for the Lie splitting, the algorithm would be the following. First, let the initial time be $t = 0$ and fix a $T$ be the final time we wish the system to reach.

1. Freeze sub-lattices belonging to $G_1$, apply KMC to each sub-lattice belonging to $G_2$ until a time $\Delta t$ is reached for all of them.

2. Since the two groups are not independent, communicate the changes in boundaries of sub-lattices in $G_2$ to the corresponding sub-lattices in $G_1$.

3. Freeze sub-lattices in $G_2$, apply KMC to each sub-lattice in $G_1$ until a time $\Delta t$ is reached for all of them.

4. Communicate changes in the boundaries of sub-lattices of $G_1$ to the corresponding sub-lattices in $G_2$.

5. Set time $t \leftarrow t + \Delta t$. Return to step 1 and repeat until $t = T$ is reached.

The significance of such an algorithm rests on the fact that state transitions between different sub-lattices in $G_1$ or $G_2$ are independent within the group. Thus, when the user applies KMC in step 1 or 3, it can be applied to the whole group $G_1$ or $G_2$ asynchronously, which leads to the speed-up in simulation.

# APPENDIX B

# SUPPLEMENTARY MATERIAL FOR CHAPTER 2

Let $X_n$ be a Markov process with $P$ the Markov transition kernel and $\mu$ the corresponding stationary distribution. Also,

$$p(\sigma_0, \ldots, \sigma_m) = p(X_0 = \sigma_0, \ldots, X_m = \sigma_0),$$

$\sigma_i$ being states of the process from a state space $S$. We also use the notation $\sigma_{0:m}$ for the sequence of states $\sigma_0, \ldots, \sigma_m$. In some cases those states will have to be distinct, and this will be mentioned separately when is needed.

## B.1 Connection of the Entropy Production with the Entropy Production Rate

In the main text (see Equations (2.14), (2.15)), we sketched a proof for the connection between entropy production (EP) for paths of length $m$,

$$\mathrm{EP}(P) = \sum_{\sigma_{0:m}} p(\sigma_{0:m}) \log \left( \frac{p(\sigma_{0:m})}{p(\sigma_{m:0})} \right), \tag{B.1}$$

and entropy production per unit time (or entropy production rate (EPR)),

$$\mathrm{EPR}(P) = \sum_{\sigma,\sigma'} \mu(\sigma) P(\sigma, \sigma') \log \left( \frac{P(\sigma, \sigma')}{P(\sigma', \sigma)} \right). \tag{B.2}$$

**Lemma 37.** *let $m \in \mathbb{N}$ and let $P$ be a Markov transition probability kernel, with $\mu$ being the stationary distribution that corresponds to $P$. Then,*

$$\mathrm{EP}(P) = m \cdot \sum_{\sigma_0,\sigma_1} \mu(\sigma_0) P(\sigma_0,\sigma_1) \log\left(\frac{P(\sigma_0,\sigma_1)}{P(\sigma_1,\sigma_0)}\right) = m \cdot \mathrm{EPR}(P). \tag{B.3}$$

*Proof.* By the Markov property, we can express $p(\sigma_{0:m})$, $p(\sigma_{m:0})$ with respect to $P, \mu$ :

$$\begin{aligned} p(\sigma_{0:m}) = &\quad \mu(\sigma_0) P(\sigma_0,\sigma_1)\cdots P(\sigma_{m-1},\sigma_m), \\ p(\sigma_{m:0}) = &\quad \mu(\sigma_m) P(\sigma_m,\sigma_{m-1})\cdots P(\sigma_1,\sigma_0). \end{aligned} \tag{B.4}$$

Substituting those in the definition of the EP in Equation (B.1), we get

$$\begin{aligned} &\sum_{\sigma_{0:m}} \mu(\sigma_0) P(\sigma_0,\sigma_1)\cdots P(\sigma_{m-1},\sigma_m) \log\left(\frac{\mu(\sigma_0) P(\sigma_0,\sigma_1)\cdots P(\sigma_{m-1},\sigma_m)}{\mu(\sigma_m) P(\sigma_m,\sigma_{m-1})\cdots P(\sigma_1,\sigma_0)}\right) \\ &= \sum_{\sigma_{0:m}} \mu(\sigma_0) P(\sigma_0,\sigma_1)\cdots P(\sigma_{m-1},\sigma_m) \log\left(\frac{\mu(\sigma_0)}{\mu(\sigma_m)}\right) \qquad \text{(B.5)} \\ &+ \sum_{\sigma_{0:m}} \mu(\sigma_0) P(\sigma_0,\sigma_1)\cdots P(\sigma_{m-1},\sigma_m) \sum_{k=1}^{m} \log\left(\frac{P(\sigma_{k-1},\sigma_k)}{P(\sigma_k,\sigma_{k-1})}\right). \qquad \text{(B.6)} \end{aligned}$$

First, we shall show that Equation (B.5) is equal to zero. We can write it as

$$\sum_{\sigma_{0:m}} \mu(\sigma_0) \log(\mu(\sigma_0)) P(\sigma_0,\sigma_1)\cdots P(\sigma_{m-1},\sigma_m) \tag{B.7}$$

$$- \sum_{\sigma_{0:m}} \mu(\sigma_0) P(\sigma_0,\sigma_1)\cdots P(\sigma_{m-1},\sigma_m) \log(\mu(\sigma_m)). \tag{B.8}$$

Now, for the first sum, we can repeatedly use that

$$\sum_{\sigma'} P(\sigma,\sigma') = 1 \tag{B.9}$$

for all states $\sigma$, which results to Equation (B.7) being reduced to

$$\sum_{\sigma_0} \mu(\sigma_0) \log(\mu(\sigma_0)).$$

For the part in (B.8), since $\mu$ is the stationary distribution associated with $P$, we have that for any state $\sigma'$,

$$\mu(\sigma') = \sum_{\sigma} \mu(\sigma) P(\sigma, \sigma'). \tag{B.10}$$

Using the property in (B.10) repeatedly on Equation (B.8), we get that it is equal to (B.7), which gives the equality of the first sum in the right-hand side of Equation (B.5) to zero. Next, we need to account for (B.6), which we write as

$$\sum_{k=1}^{m} \sum_{\sigma_{0:m}} \mu(\sigma_0) P(\sigma_0, \sigma_1) \cdots P(\sigma_{m-1}, \sigma_m) \log \left( \frac{P(\sigma_{k-1}, \sigma_k)}{P(\sigma_k, \sigma_{k-1})} \right). \tag{B.11}$$

For $k = 1$, and by using property (B.9), we get

$$\sum_{\sigma_{0:1}} \mu(\sigma_0) P(\sigma_0, \sigma_1) \log \left( \frac{P(\sigma_0, \sigma_1)}{P(\sigma_1, \sigma_0)} \right). \tag{B.12}$$

For any other $k$ in Equation (B.11), we can use Equation (B.10) to show that all terms are equal to (B.12). Since we have $m$ of those, this proves the result. □

The technique with which we showed that the term in (B.5) is equal to zero is a generalization of the one we used in the proof of Theorem 22 in the main text (see from Equation (2.31) in the main text and below).

## B.2 Connectivity and Markov generators

We remind here that $L$ is a generator of a Markov process $X_n$, $L^k$ represents the result of $k$ compositions of $L$. $d$ is the geodesic distance between states, defined with respect to the transition rates of the exact Markov process with transition probabilities $P_{\Delta t}$:

$$d(\sigma,\sigma') := \begin{cases} \min\{|\vec{z}| : \vec{z} \in \mathrm{Path}(\sigma \to \sigma')\}, & \mathrm{Path}(\sigma \to \sigma') \neq \emptyset, \\ \infty, & \mathrm{Path}(\sigma \to \sigma') = \emptyset. \end{cases} \tag{B.13}$$

In (B.13), $|\vec{z}|$ is the length of a path from $\sigma$ to $\sigma'$ and $\mathrm{Path}(\sigma \to \sigma')$ corresponds to the set of all such possible paths connecting $\sigma$ and $\sigma'$.

In the proof of Lemma 24 in the main text we used that if we have two states $\sigma, \sigma'$ with $d(\sigma,\sigma') = k$, then $L^k(\sigma,\sigma') > 0$. This is a consequence of a specific representation that $L^k(\sigma,\sigma')$ has when the states $\sigma$ and $\sigma'$ are $k$ steps apart.

**Lemma 38.** *Let $\sigma, \sigma' \in S$ and let $L$ be the generator of the Markov process. Then*

$$d(\sigma,\sigma') = k \Rightarrow L^k(\sigma,\sigma') = \sum_{z_{1:k-1}} q(\sigma, z_1) \dots q(z_{k-1}, \sigma').$$

Note the notation $z_{1:n-1} = (z_1, \dots, z_{n-1})$ for a path of states of length $n-1$. Here we assume that $\sigma, z_1, \dots, z_{n-1}, \sigma'$ are distinct states, so that the path from $\sigma$ to $\sigma'$ is of length $n$.

*Proof.* The result is immediate for $k = 0$ or $k = 1$, as $L^0(\sigma,\sigma) = \delta_\sigma(\sigma) = 1$ and $L(\sigma,\sigma') = q(\sigma,\sigma')$, since there is only one path between $\sigma$ and $\sigma'$. Let us assume that this fact holds for $k = n$. That is, for states such that $d(\sigma,\sigma') = n$,

$$L^n(\sigma,\sigma') = \sum_{z_{1:n-1}} q(\sigma, z_1) \dots q(z_{n-1}, \sigma'). \tag{B.14}$$

Note that in Equation (B.14), we have a sum that contains all paths of length $n$ connecting $\sigma$ to $\sigma'$. As the states in the sum are distinct, the product $q(\sigma, z_1) \dots q(z_{n-1}, \sigma')$ is always non-negative. In fact, an implication of representation (B.14) for $L^n(\sigma,\sigma')$

is that $L^n$ is positive when the states $\sigma$ and $\sigma'$ are $n$ steps apart. We demonstrate this now as it will be useful for the rest of the proof. Consider a path of states of length $n$ from $b_0 = \sigma$ to $b_n = \sigma'$, $(b_0, b_1, \ldots, b_{n-1}, b_n)$, where the $z_i$ are all distinct states. Then, as that sequence of states is a path, we have $q(b_i, b_{i+1}) > 0$ for $i = 0, \ldots, n-1$. However, this path is also contained in the sum in Equation (B.14). Therefore, we have

$$L^n(\sigma, \sigma') = \sum_{z_{1:n-1}} q(\sigma, z_1) \ldots q(z_{n-1}, \sigma') \geq q(\sigma, b_1) \ldots q(b_{n-1}, \sigma') > 0.$$

We will now show the result for $d(\sigma, \sigma') = n + 1$. Since $L^{n+1}$ is $L$ after $n + 1$ compositions, we can write

$$L^{n+1}(\sigma, \sigma') = L[L^n[\delta_{\sigma'}]](\sigma). \tag{B.15}$$

Then, by the definition of the generator $L$,

$$\begin{aligned} L[L^n[\delta_{\sigma'}]](\sigma) &= \sum_z q(\sigma, z) \left(L^n[\delta_{\sigma'}](z) - L^n[\delta_{\sigma'}](\sigma)\right) \\ &= \sum_z q(\sigma, z) L^n[\delta_{\sigma'}](z) \end{aligned} \tag{B.16}$$

In (B.16), we used that $d(\sigma, \sigma') = n + 1 \Rightarrow L^n[\delta_{\sigma'}](\sigma) = 0$. This is true by the induction hypothesis we made in Equation (B.14). If $q(\sigma, z) = 0$, the corresponding terms are also zero, so let $z$ be a state such that $q(\sigma, z) > 0$. As we argued above, due to the representation in (B.14) $L^n(z, \sigma') > 0$. Thus, we will now show that

$$n \leq d(z, \sigma') \leq n + 2.$$

For the upper bound, we apply the triangle inequality. To get the lower, if $d(\sigma, z) = 1$ and $d(z, \sigma')$ is lower or equal to $n - 1$, then by following the path $\sigma \to z \to \sigma'$, we get

a new path between $\sigma$ and $\sigma'$ with at most $n$ steps. This contradicts that $d(\sigma, \sigma')$ is the minimum number of steps to get from $\sigma$ to $\sigma'$, as we have already assumed that $d(\sigma, \sigma') = n + 1$.

Now, since $d(\sigma, \sigma') > n \Rightarrow L^n[\delta_{\sigma'}](\sigma) = 0$, we get that only the pairs of states $(z, \sigma')$ such that $d(z, \sigma') = n$ lead to potential non-zero terms for the sum in Equation (B.16). Therefore, if we assume $d(z, \sigma') = n$, and by using the induction step in Equation (B.16), we have

$$L^{n+1}(\sigma, \sigma') = L[L^n[\delta_{\sigma'}]](\sigma) = \sum_{z, z_{1:n-1}} q(\sigma, z) q(\sigma, z_1) \dots q(z_{n-1}, \sigma'). \tag{B.17}$$

□

While proving Lemma 38, we also demonstrated that compositions of the generators are always positive on certain pairs of states.

**Lemma 39.** *Let $\sigma, \sigma'$ be states such that $d(\sigma, \sigma') = k$. Then $L^k(\sigma, \sigma') > 0$.*

*Proof.* This is a corollary of Lemma 38. □

## B.3 Highest-order coefficients for Lie and Strang operator splitting schemes

Let $L$ be a bounded operator, which allows us to represent the semigroup $e^{Lt}$ via a power series expansion. We shall use the notation $L(\sigma, \sigma') := L[\delta'_\sigma](\sigma)$, with which we have

$$P_t(\sigma, \sigma') = e^{Lt} \delta_{\sigma'}(\sigma) = \sum_{k=0}^{\infty} \frac{L^k(\sigma, \sigma')}{k!} t^k. \tag{B.18}$$

We assume that we can write an expansion for $Q_{\Delta t}$ too by representing each semigroup in Equation (2.3) and Equation (2.4) by its series and then multiplying out. By this process, we get

$$Q_{\Delta t}(\sigma, \sigma') = \sum_{k=0}^{\infty} \frac{L_Q^k(\sigma, \sigma')}{k!} \Delta t^k, \tag{B.19}$$

where $L_Q^k$ represents the terms of order $k$ in the expansion of $Q_{\Delta t}$. For example, for the Lie splitting, $L^0_{\text{Lie}} = I, L^1_{\text{Lie}} = L, L^2_{Lie} = (L_1^2 + L_2^2 + 2L_1L_2)$. In general, the exact form of $L_Q^k$ can be computed by using the BCH formula. This notation is picked for clarity and does not imply that $L_Q$ is a generator of a Markov process. As such, $L_Q^k$ does not equal $k$ compositions of $L_Q$, except if $k < p$, $p$ being the order of the local error for the operator splitting scheme.

Lemma 40 and Lemma 41 demonstrate the form of the highest-order coefficients of the RER and the discrepancy for the Lie and Strang schemes in the case that $d(\sigma, \sigma') = 1$ implies $\sigma' = \sigma^x$ for some $x$ in the lattice. This includes the adsorption/desorption systems, an example of which was demonstrated in Section 2.3.

**Lemma 40.** *Under the assumptions of Lemma 23, if $A_{\text{Lie}}(A_{\text{Str}})$ is the highest order coefficient of the RER for the Lie (Strang) splitting, then*

$$\begin{aligned}
&A_{\text{Lie}} = \mathbb{E}_{\mu_{\text{Lie}}(\sigma)}\left[\sum_{x,y\in\Lambda} F_{\text{Lie}}(\sigma, \sigma^{x,y})\right] = \sum_{\sigma} \mu_{\text{Lie}}(\sigma) \sum_{x,y\in\Lambda} F_{\text{Lie}}(\sigma, \sigma^{x,y}),\\
&F_{\text{Lie}}(\sigma, \sigma') := C_{\text{Lie}}(\sigma, \sigma')M_{\text{Lie}}(\sigma, \sigma') - 2L^2_{\text{Lie}}[\delta_{\sigma'}](\sigma)(\text{arctanh}(M_{\text{Lie}}(\sigma, \sigma')) - M_{\text{Lie}}(\sigma, \sigma')),\\
&M_{\text{Lie}}(\sigma, \sigma') := C_{\text{Lie}}(\sigma, \sigma')/(L^2_{\text{Lie}}[\delta_{\sigma'}(\sigma)] + C_{\text{Lie}}(\sigma, \sigma'))
\end{aligned} \tag{B.20}$$

*$C_{\text{Lie}}$ stands for the commutator of the Lie scheme, $C_{\text{Lie}}(\sigma, \sigma') = [L_1, L_2]\delta_{\sigma'}(\sigma)$*

*and, for the Strang splitting,*

$$A_{\mathrm{Str}} = \mathbb{E}_{\mu_{\mathrm{Str}}(\sigma)}\left[\sum_{x,y,z\in\Lambda} F_{\mathrm{Str}}(\sigma,\sigma^{x,y,z})\right] = \sum_{\sigma}\mu_{\mathrm{Str}}(\sigma)\sum_{x,y,z\in\Lambda} F_{\mathrm{Str}}(\sigma,\sigma^{x,y,z}),$$

$$F_{\mathrm{Str}}(\sigma,\sigma') := C_{\mathrm{Str}}(\sigma,\sigma')M_{\mathrm{Str}}(\sigma,\sigma') - 2L^3_{\mathrm{Str}}[\delta_{\sigma'}](\sigma)(\mathrm{arctanh}(M_{\mathrm{Str}}(\sigma,\sigma')) - M_{\mathrm{Str}}(\sigma,\sigma')),$$

$$M_{\mathrm{Str}}(\sigma,\sigma') := C_{\mathrm{Str}}(\sigma,\sigma')/(L^3_{\mathrm{Str}}[\delta_{\sigma'}](\sigma) + C_{\mathrm{Str}}(\sigma,\sigma')). \tag{B.21}$$

*Proof.* See proof of Theorem 5.2 in [26]. □

Similarly, from the proof of Lemma 24, Section 2.4, and specifically Equation (2.38), we can write down the highest-order coefficient for the discrepancy.

**Lemma 41.** *Under the assumptions of Theorem 24, if $D_{\mathrm{Lie}}(D_{\mathrm{Str}})$ is the highest order coefficient of $I$ for the Lie (Strang) splitting, then*

$$\begin{aligned} D_{\mathrm{Lie}} = &\sum_{\sigma}\mu_{\mathrm{Lie}}(\sigma)\sum_{x\in\Lambda}\frac{q(\sigma,\sigma^x)}{q(\sigma^x,\sigma)}C_{\mathrm{Lie}}(\sigma^x,\sigma) \\ &+ \sum_{x,y\in\Lambda}\mu_{\mathrm{Lie}}(\sigma)L^2_{\mathrm{Lie}}(\sigma,\sigma^{x,y})\mathrm{atanh}(M_{\mathrm{Lie}}(\sigma^{x,y},\sigma)). \end{aligned} \tag{B.22}$$

*and*

$$\begin{aligned} D_{\mathrm{Str}} = &\sum_{\sigma}\mu_{\mathrm{Str}}(\sigma)\left(\sum_{x\in\Lambda}\frac{q(\sigma,\sigma^x)}{q(\sigma^x,\sigma)}C_{\mathrm{Str}}(\sigma^x,\sigma) + \sum_{x,y\in\Lambda}\frac{L^2(\sigma,\sigma^{x,y})}{L^2(\sigma^{x,y},\sigma)}C_{\mathrm{Str}}(\sigma^{x,y},\sigma)\right. \\ &\left. + \sum_{x,y,z\in\Lambda} L^3_{\mathrm{Str}}(\sigma,\sigma^{x,y,z})\frac{1}{3}\mathrm{atanh}(M_{\mathrm{Str}}(\sigma^{x,y,z},\sigma))\right). \end{aligned} \tag{B.23}$$

## B.4 Adsorption/Desorption Example

Here we include the setup for the adsorption/desorption example we simulated with the help of SPPARKS.

Let $\Lambda \subset \mathbb{Z}^2$ be a bounded, two-dimensional integer lattice with dimensions $N \times N$. To every lattice site $x$ corresponds a spin variable $\sigma(x)$, $\sigma(x) \in \Sigma = \{0, 1\}$, where $\sigma(x) = 0$ denotes that site $x$ is empty and $\sigma(x) = 1$ that the site is occupied by some

particle. The transition rates will correspond to single spin-flip Arrhenius dynamics. If we fix a state $\sigma \in S$ and a lattice site $x \in \Lambda$, then the transition rates $q$ are defined by

$$q(\sigma, \sigma^x) = q(x, \sigma) = c_1(1 - \sigma(x)) + c_2\sigma(x)e^{-\beta U(x)}, \tag{B.24}$$

$$U(x, \sigma) = J_0 \sum_{y \in \Omega_x} \sigma(y) + h. \tag{B.25}$$

The constants, $c_1, c_2, \beta, J_0, h$, can be tuned to generate different dynamics. $\sigma^x$ is the resulting state after starting with $\sigma$ and changing $\sigma(x)$ to $1 - \sigma(x)$. $\Omega_x$ represents the set of lattice sites that are neighbors of $x$. For this model, $\Omega_x$ will just be the nearest neighbors of $x$, like in Figure 2.1. This is the kind of spatial dependence on information that allows us to use parallel KMC. The single spin-flip process, defined by the transition rates in (B.24), satisfies detailed balance and can be simulated exactly via Kinetic Monte Carlo.

To produce Figures 2.3 and 2.4, we simulated an adsorption-desorption system with the Lie and Strang schemes in SPPARKS. The Lie scheme was already part of the software suite, whereas the Strang scheme was implemented by the authors. In order to estimate the EPR by using the expressions in Lemmas 40 and 41, we selected a splitting time-step $\Delta t = 0.001$ and simulated the process in time for $T = 100, N = 100$ while simultaneously tracking the mean coverage of the lattice (to assess equilibration of the system). Then the approximation to the EPR for the $\Delta t$ considered is given by $(A + D)\Delta t^{p-1}$.

The simulations were carried out with one processor running the parallel algorithm, although the calculation of the coefficients in Lemmas 40 and 41 can be scattered to the processes used during a parallel simulation. One of the points of the main text is that those quantities can be thought as diagnostics for the schemes and,

as such, estimated from the simulation of systems with smaller size than the target system.

## B.5 Estimators of the EPR for a Diffusion Process

To show how the calculation of the estimators would change under a different model, we shall now demonstrate the case of a diffusion process. Let us assume that it is modeled by the set of transition rates

$$q(x, y, \sigma) = p(x, y)\sigma(x)(1 - \sigma(y)),\ x, y \in \Lambda, \tag{B.26}$$

for some state $\sigma$. At each time step, the system can swap the values between two lattice sites, $x, y$. $p(x, y)$ corresponds to some decaying potential that captures the distance a particle can travel. For instance, for nearest neighbor jumps, we would have $p(x, y) = 1/4$ if $|x - y| = 1$, and $p(x, y) = 0$ otherwise. Note that the transition rates $q$ are zero if the origin site $x$ is empty or if the target site $y$ is occupied.

We focus on the case of computing the discrepancy term, $I$, for the Lie splitting with a splitting of the generator $L$ into $L_1 + L_2$. Nevertheless, this example will also be instructive for the case of the relative entropy rate and other splittings.

Theorems 23 and 24 make no assumption on the underlying model. They do however use the notion of distance between states that the transition rates define (see discussion at the beginning of Section 2.4). For this model, two states $\sigma, \sigma'$, are one jump apart if there exist distinct lattice sites $x, y$ such that $\sigma' = \sigma^{x,y}$, and two jumps apart if there exist distinct $x, y, z, w$ such that $\sigma' = \sigma^{x,y,z,w}$. The notation $\sigma^{x,y,z,w}$ denotes the resulting state after starting with a state $\sigma$ and carrying out spin-flips at the lattice locations $x, y, z, w$.

After computing the corresponding commutator, $C_{\text{Lie}}(\sigma, \sigma') = [L_1, L_2]\delta_{\sigma'}(\sigma)$, and $L^2_{\text{Lie}}$, we can write the exact formula for the highest order coefficient for the Lie splitting as

$$\begin{aligned} D_{\text{Lie}} = &\sum_{\sigma} \mu_{\text{Lie}}(\sigma) \sum_{x,y \in \Lambda} \frac{q(\sigma, \sigma^{x,y})}{q(\sigma^{x,y}, \sigma)} C_{\text{Lie}}(\sigma^{x,y}, \sigma) \\ &+ \sum_{x,y,z,w \in \Lambda} \mu_{\text{Lie}}(\sigma) L^2_{\text{Lie}}(\sigma, \sigma^{x,y,z,w}) \text{atanh}(M_{\text{Lie}}(\sigma^{x,y,z,w}, \sigma)). \end{aligned} \tag{B.27}$$

Since the commutator $C_{\text{Lie}}$ is zero for all choices of lattice sites but those at the boundaries between sub-lattices, $\partial\Lambda$, Equation (B.27) is actually

$$\begin{aligned} D_{\text{Lie}} = &\sum_{\sigma} \mu_{\text{Lie}}(\sigma) \sum_{x,y \in \partial\Lambda} \frac{q(\sigma, \sigma^{x,y})}{q(\sigma^{x,y}, \sigma)} C_{\text{Lie}}(\sigma^{x,y}, \sigma) \\ &+ \sum_{x,y,z,w \in \partial\Lambda} \mu_{\text{Lie}}(\sigma) L^2_{\text{Lie}}(\sigma, \sigma^{x,y,z,w}) \text{atanh}(M_{\text{Lie}}(\sigma^{x,y,z,w}, \sigma)). \end{aligned} \tag{B.28}$$

Therefore, for nearest neighbor interactions in a square $N \times N$ lattice, the coefficient in (B.28) has cost of computation $O(N^2)$. Note that the difference in scaling of the cost is because of the underlying diffusion dynamics and which imply that $d(\sigma, \sigma') = 1$, that is, the states the system can reach in one step from $\sigma$, are precisely $\sigma' = \sigma^{x,y}$ for $x, y$ distinct lattice sites. However, estimating coefficient (B.28) is more of a diagnostic that does not have to be computed while simulating the large system, which is why we normalize coefficients by their scaling while estimating.

# BIBLIOGRAPHY


[1] Abdulle, Assyr, Vilmart, Gilles, and Zygalakis, Konstantinos C. Long time accuracy of Lie–Trotter splitting methods for Langevin dynamics. *SIAM Journal on Numerical Analysis 53*, 1 (2015), 1–16.

[2] Akaike, Hirotogu. Information theory and an extension of the maximum likelihood principle. In *Selected Papers of Hirotugu Akaike*, Emanuel Parzen, Kunio Tanabe, and Genshiro Kitagawa, Eds., Springer Series in Statistics. Springer New York, 1998, pp. 199–213.

[3] Akaike, Hirotugu. A new look at the Bayes procedure. *Biometrika 65*, 1 (1978), 53–59.

[4] Arampatzis, Giorgos, Katsoulakis, Markos A., and Plecháč, Petr. Parallelization, processor communication and error analysis in lattice kinetic Monte Carlo. *SIAM Journal on Numerical Analysis 52*, 3 (2014), 1156–1182.

[5] Arampatzis, Giorgos, Katsoulakis, Markos A., Plecháč, Petr, Taufer, Michela, and Xu, Lifan. Hierarchical fractional-step approximations and parallel kinetic Monte Carlo algorithms. *J. Comput. Phys. 231*, 23 (Oct. 2012), 7795–7814.

[6] Bayati, Basil S. Fractional diffusion-reaction stochastic simulations. *The Journal of Chemical Physics 138*, 10 (2013).

[7] Bilionis, I, and Koutsourelakis, Phaedon-Stelios. Free energy computations by minimization of kullback–leibler divergence: An efficient adaptive biasing potential method for sparse representations. *Journal of Computational Physics 231*, 9 (2012), 3849–3870.

[8] Bilionis, Ilias, and Zabaras, Nicholas. A stochastic optimization approach to coarse-graining using a relative-entropy framework. *The Journal of chemical physics 138*, 4 (2013), 044313.

[9] Bishop, C. Pattern recognition and machine learning (information science and statistics), 1st edition. 2006. corr. 2nd printing edition, 2007.

[10] Blei, David M, Jordan, Michael I, et al. Variational inference for dirichlet process mixtures. *Bayesian analysis 1*, 1 (2006), 121–143.

[11] Boucheron, Stéphane, Lugosi, Gábor, Massart, Pacal, et al. On concentration of self-bounding functions. *Electron. J. Probab 14*, 64 (2009), 1884–1899.

[12] Boucheron, Stéphane, Lugosi, Gábor, and Massart, Pascal. A sharp concentration inequality with applications. *Random Structures & Algorithms 16*, 3 (2000), 277–292.

[13] Boucheron, Stéphane, Lugosi, Gábor, and Massart, Pascal. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

[14] Chaimovich, A., and Shell, M. S. Relative entropy as a universal metric for multiscale errors. *Phys. Rev. E 81*, 6 (2010), 060104.

[15] Chareka, Patrick, Chareka, Ottilia, and Kennendy, Sarah. Locally sub-gaussian random variables and the strong law of large numbers. *Atlantic Electronic Journal of Mathematics 1*, 1 (2006), 75–81.

[16] Chowdhary, Kamaljit, and Dupuis, Paul. Distinguishing and integrating aleatoric and epistemic variation in uncertainty quantification. *ESAIM: Mathematical Modelling and Numerical Analysis 47*, 03 (2013), 635–662.

[17] Chowdhary, Kamaljit, and Dupuis, Paul. Distinguishing and integrating aleatoric and epistemic variation in uncertainty quantification. *ESAIM: M2AN 47*, 3 (2013), 635–662.

[18] Cover, Thomas M., and Thomas, Joy A. *Elements of Information Theory*. Wiley-Interscience, New York, NY, USA, 1991.

[19] Dashti, Masoumeh, and Stuart, Andrew M. The bayesian approach to inverse problems. *arXiv preprint arXiv:1302.6989* (2013).

[20] Dembo, Amir, and Zeitouni, Ofer. Large deviations techniques and applications, volume 38 of stochastic modelling and applied probability, 2010.

[21] Dupuis, Paul, Katsoulakis, Markos A., Pantazis, Yannis, and Plechác, Petr. Path-space information bounds for uncertainty quantification and sensitivity analysis of stochastic dynamics. *SIAM/ASA Journal on Uncertainty Quantification 4*, 1 (2016), 80–111.

[22] Dupuis, Paul, Katsoulakis, Markos A., Pantazis, Yiannis., and Plecháč, Petr. Path-space information bounds for uncertainty quantification and sensitivity analysis of stochastic dynamics. *ArXiv e-prints* (Mar. 2015).

[23] Dupuis, Paul G., and Ellis, Richard Steven. *A weak convergence approach to the theory of large deviations*. Wiley series in probability and statistics. Wiley, New York, 1997. A Wiley-Interscience publication.

[24] Engblom, Stefan, Ferm, Lars, Hellander, Andreas, and Ltstedt, Per. Simulation of stochastic reaction-diffusion processes on unstructured meshes. *SIAM Journal on Scientific Computing 31*, 3 (2009), 1774–1797.

[25] Gallavotti, Giovanni, and Cohen, EGD. Dynamical ensembles in stationary states. *Journal of Statistical Physics 80*, 5-6 (1995), 931–970.

[26] Gourgoulias, Konstantinos, Katsoulakis, Markos A., and Rey-Bellet, Luc. Information metrics for long-time errors in splitting schemes for stochastic dynamics and parallel kMC. *pre-print* (2015). arXiv:1511.08240 [math.NA].

[27] Gourgoulias, Konstantinos, Katsoulakis, Markos A., and Rey-Bellet, Luc. Information metrics for long-time errors in splitting schemes for stochastic dynamics and parallel kinetic monte carlo. *SIAM Journal on Scientific Computing 38*, 6 (2016), A3808–A3832.

[28] Gourgoulias, Konstantinos, Katsoulakis, Markos A., and Rey-Bellet, Luc. Information criteria for quantifying loss of reversibility in parallelized KMC. *Journal of Computational Physics 328* (2017), 438 – 454.

[29] Hansen, Eskil, and Ostermann, Alexander. Exponential splitting for unbounded operators. *Mathematics of computation 78*, 267 (2009), 1485–1496.

[30] Hellander, Andreas, Lawson, Michael J., Drawert, Brian, and Petzold, Linda. Local error estimates for adaptive simulation of the reaction-diffusion master equation via operator splitting. *J. Comput. Phys. 266* (June 2014), 89–100.

[31] Hoeffding, Wassily. Probability inequalities for sums of bounded random variables. *Journal of the American statistical association 58*, 301 (1963), 13–30.

[32] Hoffman, Matthew D, Blei, David M, Wang, Chong, and Paisley, John William. Stochastic variational inference. *Journal of Machine Learning Research 14*, 1 (2013), 1303–1347.

[33] Jahnke, Tobias, and Altntan, Derya. Efficient simulation of discrete stochastic reaction systems with a splitting method. *BIT Numerical Mathematics 50*, 4 (2010), 797–822.

[34] Kalligiannaki, Evangelia, Harmandaris, Vagelis, Katsoulakis, Markos A., and Plecháč, Petr. The geometry of generalized force matching and related information metrics in coarse-graining of molecular systems. *The Journal of Chemical Physics 143*, 8 (2015).

[35] Katsoulakis, Markos, Pantazis, Yannis, and Rey-Bellet, Luc. Measuring the irreversibility of numerical schemes for reversible stochastic differential equations. *ESAIM: Mathematical Modelling and Numerical Analysis 48* (9 2014), 1351–1379.

[36] Katsoulakis, Markos A. Rey-Bellet, Luc, and Wang, Jie. Scalable information inequalities for uncertainty quantification, 2016. arXiv:1605.04184.

[37] Katsoulakis, Markos A., Rey-Bellet, Luc, and Wang, Jie. Scalable information inequalities for uncertainty quantification. *Journal of Computational Physics* (2017), 513–545.

[38] Kelly, Frank P. *Reversibility and stochastic networks*. Cambridge University Press, 2011.

[39] Kipnis, Claude, and Landim, Claudio. *Scaling limits of interacting particle systems*. Grundlehren der mathematischen Wissenschaften. Springer, Berlin, New York, 1999. Appendix 1.

[40] Kirkpatrick, Keith. Battling algorithmic bias: how do we ensure algorithms treat us fairly? *Communications of the ACM 59*, 10 (2016), 16–17.

[41] Kurchan, Jorge. Fluctuation theorem for stochastic dynamics. *Journal of Physics A: Mathematical and General 31*, 16 (1998), 3719.

[42] Lebowitz, Joel L., and Spohn, Herbert. A Gallavotti–Cohen-type symmetry in the large deviation functional for stochastic dynamics. *Journal of Statistical Physics 95*, 1 (1999), 333–365.

[43] Lehmann, Erich Leo, and Casella, George. *Theory of point estimation*. Springer Science & Business Media, 2006.

[44] Leimkuhler, Ben, and Matthews, Charles. *Molecular Dynamics: with deterministic and stochastic numerical methods*, vol. 39. Springer, 2015.

[45] Leimkuhler, Benedict, Matthews, Charles, and Stoltz, Gabriel. The computation of averages from equilibrium and nonequilibrium Langevin molecular dynamics. *IMA Journal of Numerical Analysis* (2015).

[46] Lelièvre, Tony, Stoltz, Gabriel, and Rousset, Mathias. *Free energy computations: A mathematical perspective*. World Scientific, 2010.

[47] Liu, Jun S. *Monte Carlo strategies in scientific computing*. Springer Science & Business Media, 2008.

[48] MacKay, David JC. *Information theory, inference and learning algorithms*. Cambridge university press, 2003.

[49] Maes, Christian. The fluctuation theorem as a Gibbs property. *Journal of statistical physics 95*, 1-2 (1999), 367–392.

[50] Maes, Christian, Redig, Frank, and Moffaert, Annelies Van. On the definition of entropy production, via examples. *Journal of Mathematical Physics 41*, 3 (2000), 1528–1554.

[51] Mattingly, J.C., Stuart, A.M., and Tretyakov, M. Convergence of numerical time-averaging and stationary measures via the poisson equation. *SIAM Journal of Numerical Analysis 48* (2010), 552–577.

[52] McLachlan, Robert I, and Quispel, G Reinout W. Splitting methods. *Acta Numerica 11* (2002), 341–434.

[53] Nilmeier, Jerome P., and Marian, Jaime. A rigorous sequential update strategy for parallel kinetic Monte Carlo simulation. *Computer Physics Communications 185*, 10 (2014), 2479 – 2486.

[54] Pazy, Amnon. *Semigroups of Linear Operators and Applications to Partial Differential Equations*. Springer New York, 1983.

[55] Pinski, FJ, Simpson, Gideon, Stuart, AM, and Weber, Hendrik. Kullback–leibler approximation for probability measures on infinite dimensional spaces. *SIAM Journal on Mathematical Analysis 47*, 6 (2015), 4091–4122.

[56] Plimpton, Steve, Battaile, Corbett, Chandross, Mike, Holm, Liz, Thompson, Aidan, Tikare, Veena, Wagner, Greg, Webb, E, Zhou, X, Cardona, C Garcia, et al. Crossing the mesoscale no-mans land via parallel kinetic Monte Carlo. *Sandia Report SAND2009-6226* (2009).

[57] Raginsky, Maxim, Sason, Igal, et al. Concentration of measure inequalities in information theory, communications, and coding. *Foundations and Trends® in Communications and Information Theory 10*, 1-2 (2013), 1–246.

[58] Rudzinski, [J. F., and Noid, W. G. Coarse-graining, entropy, forces and structures. *J. Chem. Phys. 135*, 21 (2011).

[59] Talay, Denis, and Tubaro, Luciano. Expansion of the global error for numerical schemes solving stochastic differential equations. *Stochastic Analysis and Applications 8*, 4 (1990), 483–509.

[60] Trotter, Hale F. On the product of semi-groups of operators. *Proceedings of the American Mathematical Society 10*, 4 (1959), pp. 545–551.

[61] Tsybakov, Alexandre B. Introduction to nonparametric estimation. revised and extended from the 2004 french original. translated by vladimir zaiats, 2009.

[62] Wainwright, JM. High-dimensional statistics: A non-asymptotic viewpoint. *preparation. University of California, Berkeley* (2015).

[63] Wainwright, Martin J, Jordan, Michael I, et al. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning 1*, 1–2 (2008), 1–305.